The Timing and Coordination of Turn-Taking

Matthew Christopher Bull

Doctor of Philosophy
University of Edinburgh
1998

I declare that this thesis has been composed by myself, and that the research which is reported herein has been conducted by myself unless otherwise indicated

Abstract

The general coordination of spoken dialogue among participants in a conversation has received considerable attention, and several theories propose a mechanism by which participants in a conversation coordinate and time entries to the conversational floor. This mechanism is not trivial, because the timing of these entries can be too closely aligned to the offset of the previous turn to be a simple response to its offset.

One hypothesis which is tested here is the *Rhythmic Coordination Hypothesis* (see *English Speech Rhythm*, by E. Couper-Kuhlen, 1993). This assumes that a listener is able to perceive regular prosodic prominences in the speech of others, and can extrapolate a rhythmic beat from the sequences of prominences. Importantly, the hypothesis maintains that the listener is able to coordinate the first prominent syllable in his or her contribution with the extrapolated rhythmic beat. In effect a listener will, by default, start a contribution 'on the beat'.

A series of experiments showed little evidence in favour of this hypothesis, and indicated instead that the context in which utterances occur may play a more significant role in the timing and coordination of turn-taking. A further set of analyses investigated this, and showed that factors such as the presence or absence of game boundaries, the move category of an utterance or part of an utterance, and the presence or absence of eye contact, significantly affect the timing of turn-taking.

The pattern of results suggests that turn-taking is not a simple response mechanism. Instead it reflects factors of different kinds and reveals both feedback and planning functions. This supports the notion that conversation is an interactive process in which the participants are involved in the problem of coordinating their respective signals well enough to be mutually intelligible, yet within certain time constraints. These constraints are set up both by the restrictions of planning and processing time, and also by social factors.

Acknowledgements

I would like to thank those members of the HCRC Dialogue Working Group who have provided me with feedback and suggestions on my research in general, and on the Map Task and its design in specific. I must also thank Matthew Aylett, who designed the Automatic Stress Assignment model used in this thesis. Special thanks are due to my supervisor Ellen Bard, who frequently went out of her way to give me every help possible, and whose seemingly boundless energy and enthusiasm for my research was invaluable. I thank also my good friends Kim Hardie and Iain Lang (as well as all my other friends) for lending me moral support and a listening ear, and for helping me put the whole thing into perspective.

Contents

1. Introduction

It is one of the remarkable facts of conversation that several people are able to coordinate their activities in a meaningful way, very often without their being consciously aware of the process. But it is not simply that people coordinate their topics of conversation, and can reach agreement or disagreement on various subjects. The essential ingredient to coordination is *time.* It is this element with which this thesis is concerned.

Conversations in a broad sense may cover communication between people in the form of letters, emails, and drawings, where a considerable amount of time elapses between a 'statement' and a 'reply'. But these media do not form the *basic* form of communication between people. Rather, it is *face-to-face* interaction which is the basic form of communication. One may describe writing in terms relative to face-to-face interaction, but not vice versa.

Time is a scarce resource in interaction. This is certainly true in conversation, where parties generally try to waste as little time as possible in coordinating their contributions. Very often, a person will start to speak before the other has finished - either accidentally or deliberately. Or there may be lapses, where no one is speaking. Whichever of these cases happens, the participants will be able to read some significance into them. Overlap, interruption, and lapses may therefore be regarded as signals. However, this leaves out a very real need to account for cross-cultural differences. I am not aware of any research which has compared the timing of turn-taking in different cultures in any rigorous or systematic way. Certainly, the research reported on in this thesis makes no attempt at a cross-cultural study. For the moment, therefore, cross-cultural factors can only be guessed at, and the most we can say is that one person will typically attempt to start speaking as soon as possible after another person has finished speaking, where the definition of 'as soon as possible' may vary according to the culture.

The precision of the coordination of timing, and people's sensitivities to timing, are also remarkable. It has long been recognized in the literature that a listener must be able to predict, or at least make a good guess of, the end of a speaker's utterance. This is because people are able to respond to utterances either before they are completely finished, or so soon after their end, that it would be impossible for them to react quickly enough if they were to wait for the end of the utterance, and only then plan and produce a response. Quite how a listener is able to predict an end-point of an utterance is not known for certain, although there is good evidence that syntactic, pragmatic, and intonational information are used.

Two basic approaches have been employed to explain coordination in conversation. One approach arises from a psychological tradition (e.g. Duncan, 1972), and assumes that contributions from each speaker are accompanied by various signals. These act as cues to a listener, informing him or her when the speaker wishes to let the listener speak, or informing the listener when an appropriate opportunity to start speaking has arisen. The two speakers interact only at a simple level, where A passes a signal to B, and B passes a signal back. The other strategy is founded in sociological treatments, and stresses a more complex interactive nature of conversation than the cue-based approach (e.g. Sacks et al., 1974). Each person does not simply receive and pass a signal. Instead, conversation is coordinated such that the speaker may offer the 'conversational floor' (in other words, a chance to speak) to one of the listeners, or one of the listeners may choose to take the conversational floor at strategic moments. If the floor is not taken, then as soon as the speaker realises this he or she may choose to continue. This interactive approach is governed by a set of 'rules' of conversational organisation.

Each of these approaches has advantages and disadvantages. The cue-based theory has the primary limitation that it ignores the interactive and social aspects of conversation. The interactive theory has no adequate way of accounting directly for the ability of a listener to predict the end of a speaker's contribution. Despite this, the interactive approach seems to have taken precedence in conversation research over the last twenty years.

There have been attempts to take the basic interactive premises, and strengthen them with a 'psychological' element. For example, part of this thesis is concerned with a hypothesis (developed by, amongst others, Elizabeth Couper-Kuhlen) which assumes the conversational 'rules' of the interactive theory, but which also claims that the temporal coordination of conversation is largely managed using the perceived rhythm of speech. According to the hypothesis, a listener perceives a rhythmic beat in the speaker's utterance, and is able to extrapolate this rhythmic beat beyond the end of the utterance. The start of his or her contribution is timed such that its first beat coincides with one of the extrapolated rhythmic beats. This need not always take place - rhythmic coordination is the unmarked case, and arhythmic coordination carries some communicative significance. Presumably, in the unmarked case of rhythmic integration, the context in which utterances take place only affects coordination inasmuch as it determines on which beat a listener may start speaking. Context, here, is used to refer to a number of factors present in a conversation, ranging from the familiarity of the participants to the possible mental processing required to understand an utterance and respond to it. The ability of listeners to predict the end of someone else's contribution would also affect the placing of the first beat.

Alternatively, it might be possible to incorporate the interactive framework with a non-rhythmic coordination mechanism. The coordination of utterances would be based purely on the context in which those utterances occur, and the accuracy of the listener's prediction of when the speaker's current contribution will finish.

But conversation is not some haphazard arrangement of signals and information. Closer inspection reveals that it is a finely coordinated process, where even slight deviations from an accepted pattern may convey signals. According to Clark (1996), conversation is essentially a coordination problem. Coordination problems involve an understanding by one person of how another might act, based on the assumption that other people will either act similarly to oneself, or will act according to various stereotypes. People use these understandings to be able to interact, and in the case of conversation, to manage a scarce resource. This scarce resource is, in effect, the ability to have an opportunity to speak. But conversation

involves more than one person's holding the conversational floor at a time. While one person is speaking, others may send out signals that let that person know that they are paying attention, or not, as the case may be. When people engage in conversation, they are doing more than just conveying facts. They are also trying to establish a certain relationship or common ground between them. It takes little insight to realize that a great deal of conversation conveys very few 'hard facts'. Much of what people say contains largely social information. The goal of conversation is for people to communicate a signal as intelligibly as is necessary for that signal to be understood by others, within the time constraints imposed by the management of a scarce resource.

Analyses of the intervals between speakers' contributions should shed light on the processes that take place when participants in a conversation coordinate their contributions. They should provide evidence of whether rhythm may be used as a coordinator, and whether any generalizations may be made of which contexts, if any, are significant factors in coordination. A problem with analysing the intervals between speakers' utterances is that this supposes that every interval is relevant. That is, it supposes that any given utterance may be thought of as being a response to some other utterance - specifically, the previous one. But a moment's reflection reveals that this need not be the case. It is possible to start a new topic of conversation, to interrupt an utterance, or to respond to an utterance which was earlier in the conversation. This problem lies at the heart of analyses of conversation, and is reflected in the general tendency for earlier theorists and researchers to view conversation rather like a tennis match, where each party takes a turn to speak, and where turns must follow one another in an ordered fashion. It is perhaps better to regard a conversation between two people as consisting of two interweaving strands of speech. Silence by one speaker does not mean that the strand stops, because silence is itself communicative. The strand simply changes. The work described in this thesis (such as work by Clark) attempts to break free from the traditional mode of thought, by chunking conversations into stretches of speech and silence by each speaker.

In this thesis, then, I set out to analyse the intervals between different speakers' utterances (inter-speaker intervals). First, I test the hypothesis that the coordination of turn-taking is achieved through the perceived rhythm of speech. The tests consist of both perceptual experiments, and analysis of intervals between prominent syllables within and across speakers' utterances in conversation taken from the HCRC Map Task Corpus. This consisted of a series of dialogues, where two people had to collaborate to achieve an overall goal. This is an important point, because task-oriented dialogue, where each speaker has a clearly-defined role as either the giver or receiver of instructions, could be quite different from other varieties of conversation, like informal chatting or highly formal interview situations.

2. Literature Review

2.1 Introduction

In this chapter I outline some basic concepts in discourse analysis. I also mention the need for a model which can describe the coordination processes that participants employ in a conversation. There are essentially two traditional models, which approach turn-taking from different academic perspectives. One is typified by a psychological, stimulus-response point of view, in which each participant in a conversation is treated as an individual who recognises the end of the other speaker's turn through a series of signals given out by that speaker (Duncan, 1972). Another model views turn-taking as an essentially socially based, interactive process, in which neither participant can be regarded separately from the other. A set of rules is used to allocate turns at holding the conversational floor (Sacks, Schegloff & Jefferson, 1974).

However, each of these approaches suffers from not incorporating aspects of the other. Neither can provide a complete account of conversation by itself. Also, both approaches wrongly, in my opinion, regard turn-taking as a well-structured process, in which each speaker takes it in turn to gain and hold the conversational floor.

I suggest that a solution is offered by an interactive model (Clark, 1996) which assumes that conversation is a form of a more general communicative process of joint actions. This model places an emphasis on treating conversation as an interactive process which is made up of individuals' acting on the actions of others, and providing continuous, direct and positive feedback for one another. Clark argues that conversation can give the illusion that it is organised around a set of ordered rules. But this orderliness emerges from basic strategies that conversationalists adopt - for example, to expect that contributions generally are made up of presentations followed by acceptances, or to expect that as far as is possible only one person should

speak at a time. The coordination of timing emerges as an element of paramount importance, because of the immediate nature of conversation. The coordination process could be carried out either using perceived rhythm (see, for example, Couper-Kuhlen, 1993), or using a simple temporal 'window' within which next-speakers must respond. For either process, a means of projecting the end of a turn is vital.

2.2 Basic Elements of Discourse Analysis

One of the basic aims of discourse analysis is to discover a set of rules which could reproduce the naturally occurring structure of discourse, and which would therefore be able to predict whether any given discourse structure was coherent or not. In this respect it is analogous to grammar, although it operates at a different level in the linguistic organisational hierarchy. This hierarchy runs from the phonological level to the non-linguistic level (see Table 1).

Table 1 - *Linguistic organisational hierarchy (Coulthard, 1977)*

phonology	*grammar*	*discourse*	*non-linguistic*
phoneme			
syllable			
	morpheme		
foot			
	word		
	group		
tone group			
	clause		
		act	
	sentence		
		move	
		exchange	
			stage
			transaction

The non-linguistic level of organisation may need some clarification. Although it has long been recognised that the study of conversation would lead to a better understanding of language (e.g. Firth, 1935[1]), there were few attempts actually to study it until studies of pragmatics and context-based speech rose in prominence in the 1960's. Mitchell (1957) provides a semantically motivated analysis of supra-sentential structure. He divides a buying/selling situation into stages, where each stage is an abstract category based on semantic criteria. So, a shop transaction may ideally consist of the following stages - 1. salutation; 2. enquiry of object of sale; 3. investigation of object of sale; 4. bargaining; 5. conclusion.

Coulthard (1977) argues that this is not a linguistic analysis at all, since 'the stages are defined and recognised by the activity that occurs within them rather than by characteristic linguistic features...' (p. 5) Though the stages are non-linguistic, their *characterisation* is linguistic. For example, cases of stage 2 (enquiry) are characterised by linguistic question-answer sequences, where one *linguistic* item (question) constrains the next speaker's choice of item (answer). Therefore, a level of supra-sentential *linguistic* structure is missing from the description.

Since it is also the linguistic *function*, not the form, of these constraints on the next speaker, the grammar level alone would not be sufficient. For example, it is quite possible to have a discourse consisting of a series of utterances which are grammatically well-formed, but which are odd from the point of view of discourse, as in 1) below.

1)
A: I feel hot today.
B: No.

(Labov, 1970)

And presumably discourse may consist of grammatical peculiarities, yet remain generally coherent.

---

[1]Cited in Coulthard (1977).

There seems to be good justification, then, for another level, founded on function, between the grammar level and the non-linguistic level - *discourse*.

If one of the main aims of discourse analysis is to discover a set of rules for well-formed discourse, it follows that a set of *discourse units* would be needed for those rules to act upon. The units are functional, and should relate in some way to formal, grammatical units. Describing this relation explicitly poses problems, since there is no direct link between form and function in language. One solution, adopted by Sacks and other conversational analysts, is to assume 'intuitively recognisable' functional units. This is not particularly satisfactory, and others have attempted to devise a set of rules which explain how a given linguistic function is linked to a given grammatical structure.

The importance of a consistent system of discourse units is highlighted by Sinclair & Coulthard (1975, 1992), who argue for the necessity of an overall descriptive framework. They claim that conversational analysts working without such a framework would fall into the trap of producing a potentially infinite number of 'data-specific descriptive categories for each new piece of text to the last syllable of recorded conversation.' (p.55) The basis of this framework is to establish what the units are, and how they relate to each other. A *rank scale* is used to link units of different types, and one may say that a unit in a rank scale consists of units at a lower level, such units consisting of units at a still lower level, and so on. The point is that using this basic framework, often 'intuitive' categories such as the utterance do not fit into a hierarchical structure. Sinclair & Coulthard (1992) note that initially they assumed two levels of analysis - the *utterance* and the *exchange*. The utterance was defined as everything said by one speaker before another begins speaking, and an exchange was defined as two or more utterances. But this analysis quickly breaks down. Example 2) appears intuitively to have three utterances, but there is some uncertainty about the third utterance.

2)

T: Can you tell me why do you eat all that food?

Yes.

P: To keep you strong.

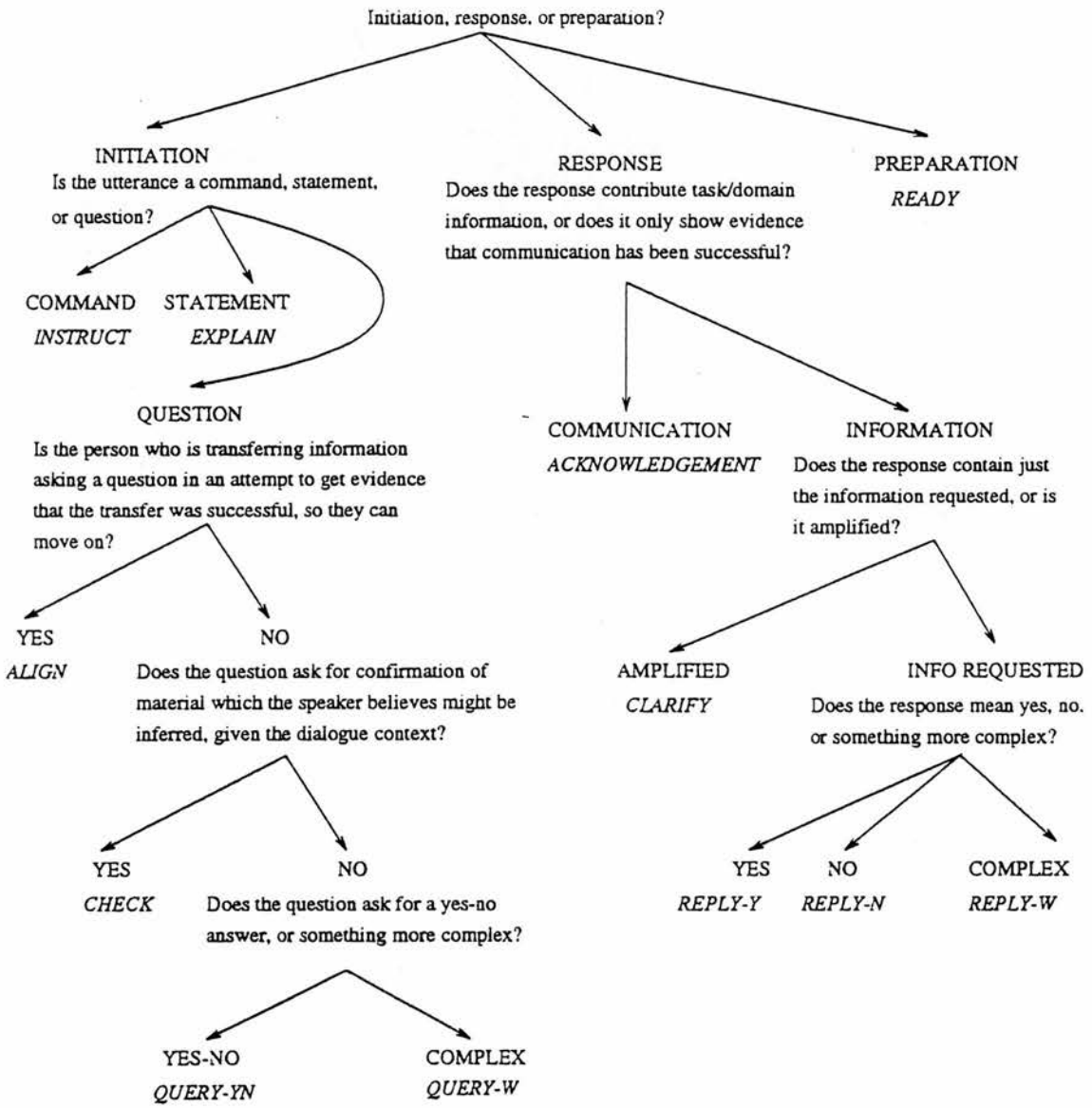
T: To keep you strong. Yes. To keep you strong. Why do you want to be strong?

(Sinclair & Coulthard, 1992. pp. 2-3)

In T's second utterance there seems to be a natural boundary before the question is asked, suggesting a level of analysis lower than the utterance. This is called the *move*. But as Sinclair & Coulthard point out, this does not then mean that there need be three levels - move, utterance, and exchange - because the utterance level now becomes superfluous. If both exchanges and moves are adopted as units of discourse they can account for the data sufficiently. Accordingly, an exchange would be defined as consisting of two or more moves under this revised definition, where each participant makes at least one move.

Sinclair & Coulthard (1992) point out that while a move is the basic free unit of discourse, it appears to have an internal structure. For example, an initiating move may have several different functional properties - *elicitation*, *direction* and *information*, where elicitation prompts a linguistic response, direction prompts a non-linguistic response, and information passes on facts, opinions, and ideas. There is, then, a need for proposing a unit below the move - the *act*. This may be described as approximating to a clause (although again it is not identical because in discourse analysis the emphasis lies with function and not form). As a result a clause may be labelled a *yes/no question* because of its form, but within a dialogic analysis its function might be *directive*. Consider, for example, the utterance 'Can you close the door?'

Carletta et al. (1996) use a system of twelve move categories, defined according to function. The distinctions they used to classify moves are shown in Figure 1.



(Carletta et al., 1996)

Figure 1 - Classification system for move coding

The move categories may be grouped into three broad types: initiating moves, responding moves, and transitional moves.

2.2.1.1 Initiating Moves

Initiating moves are those which set up the expectation of a response. Initiating moves can also act as responding moves, as in example 3) below.

3)

G: Have you got a haystack on your map?

F: Yeah

G: Right just move straight down from there, then

F: Past the blacksmith?

"Past the blacksmith?" is a *query-yn* move, which is an initiator class of move. But here it appears to act as a *response* to the previous utterance. The classification of moves into initiating and responding moves is therefore flexible.

The types of initiating move are listed below.

a) Instruct

This move acts as a command for the partner to carry out an action, except when the command is in fact a question.

b) Explain

An *explain* move states information not elicited by the partner. These moves may often result where there is some problem, and the speaker needs to explain the difference.

c) Check

A *check* move is a request for clarification about some piece of information which the checker has some reason to believe but is not certain about. For example, a *check*

may arise when a listener thinks he or she has misheard an instruction, as in example 4):

4)
Giver: Ehm, curve round slightly to your right.
Follower: To my right?

d) Align

A typical form of an *align* move is *'OK?'*, and *align* moves are used to check the attention of the partner, or that the partner has understood an instruction and is ready to continue. In this respect, an *align* move would appear to be similar in function to a *check*, the difference being that an *align* move tries to ascertain whether the listener rather than the speaker understands the situation fully.

e) Query-yn

This move type has the function of a question to which the answer would be expected to be *'yes'* or *'no'*.

f) Query-w

Query-w moves cover all types of question to which the expected answer would not be a *'yes'* or *'no'*.

2.2.1.2 Response Moves

This class of move covers moves which would generally be expected to act as a response to another move (an initiating move) in an exchange. Again, this is only a generalization, since response class moves may on occasion follow other response moves.

a) Acknowledgement

These demonstrate that the speaker has heard and understood the previous move. *Acknowledge* moves take the form of utterances like *'mmhmm'*, *'right'* or *'OK'*. They

may be considered to form the basis of backchannelled utterances (see section 2.9), and contain little or no semantic content. They have a purely discourse-related function.

b) Reply-y

A *reply-y* move is any affirmative reply to a yes-no question. They normally only appear after *query-yn*, *align*, and *check* moves.

c) Reply-n

These are negative responses to *query-yn* moves.

d) Reply-w

A *reply-w* move is any reply to a query which is not a *'yes'* or *'no'*. They typically follow *query-w* moves.

e) Clarify

A *clarify* move may be thought of as a reply plus a short *explain* move, where the explanation is not asked for explicitly. Clarification often occurs where the follower seems unsure of what to do, but there is no obvious problem. Carletta et al. (1996) claim only to use *clarify* moves where the new information is not substantial enough to be classified separately as an *explain* move.

2.2.1.3 The Ready Move

Carletta et al. (1996) classify this move type separately from initiating/responding type moves, and claim that it forms its own *transitional* move category. *Ready* moves occur between games, and often may take the form of *'OK'* or *'right'*.

2.2.2 Game Coding

According to Carletta et al. (1996), a conversational game "is a sequence of moves starting with an initiation and encompassing all moves up until that initiation's purpose is either fulfilled or abandoned." (p.11)

There are two points to consider in their game coding scheme. The first is the function of the game. In the Map Task coding scheme, a game's purpose was assumed to be the same as that of the first move in the game. So, if a game started with an *instruct* move, then it would be referred to as an *instruct* game.

The second consideration is how games are related to one another. The simplest relationships assume that once a game has been started participants work on the goal of that game until they both believe that it has been accomplished, or until they believe that the goal should be abandoned (see, for example, Houghton, 1986[2]). But as Carletta et al. (1996) point out, most natural conversation is not organized in such an orderly way:

> ...participants are free to initiate new games at any times (even while the partner is speaking), and these new games can introduce new purposes rather than serving some purpose which is already present in the dialogue. In addition, natural dialogue participants often fail to make clear to their partners what their goals are. This makes it very difficult to develop a reliable coding scheme for complete game structure. (Carletta et al., 1996 p. 11)

The game coding scheme is therefore kept relatively simple. Beginnings and ends of all games are marked. Games are also marked for being either top level or embedded. A top level game is essentially one which starts after all previous games have completely finished. An embedded game is one which starts before any previous games have ended, and is therefore a part of some higher-level goal.

---

[2] Cited in Carletta et al., 1996.

2.2.3 Transaction Coding

There is also need for a unit above games and moves. Sinclair & Coulthard (1992) labelled this category the *transaction*. This element can roughly be thought of as a stage in the discussion, and is commonly introduced by a *frame* - a word or short phrase such as 'OK', 'right', 'now' - whose function is to 'announce' a shift from one stage to another. Sinclair & Coulthard also note the presence of a special kind of statement following the frame, called the *focus*. The function of the focus statement is to introduce the content of the new stage. Presumably, the exact structure of a transaction varies greatly, and Sinclair & Coulthard's analysis was based on teacher-pupil interaction only. Nevertheless, the presence of frames and focus statements indicates structure above the exchange, something which one might have predicted intuitively.

A transaction coding scheme is given in Carletta et al. (1996). Each transaction is built up of several games, and corresponds to one step of the overall goal of the conversation. The coding scheme described in Carletta et al. is based around the Map Task Corpus (Anderson et al. 1991) - a task-oriented dialogue corpus. Since participants in a conversation had a specific overall goal to achieve, and a series of sub-goals within this, the dialogues would have been structured differently from other types of conversation (for example, informal chatting). However, the principles behind the transaction, game, and move coding adopted would remain the same.

The transaction coding scheme consists of determining how participants in a conversation divide the overall task into sub-tasks. Coding consists of the start and end points of the transaction, plus the dialogue which was used to achieve the particular sub-task. There are, however, a number of problems. First, as we saw with game coding, participants do not necessarily proceed in an orderly way from one task to the next, starting a new task only when a previous one has been completed. As problems arise, some degree of backtracking to an earlier sub-task may be necessary. Or sometimes one participant may give a brief account of a future sub-task to set a context for that task. A route giver's description which begins "So, what I'll be asking you to do is..." is an example of such an overview.

Second, participants may also review a sub-task, perhaps by describing a problem that was encountered in the conversation to be certain that it has been fully understood by both participants.

A third problem is that participants may bring in an area of discussion tangential to the overall goal of a conversation, making coding difficult. Transaction coding was therefore divided into four categories: normal, review, overview, and irrelevant.

In summary, discourse analysis is concerned with generating a set of rules which would reproduce the structure of discourse. Within a linguistic organisational hierarchy, discourse is above the grammatical level, and is concerned with the function of utterances rather than their form. The units that a hypothesised set of rules would operate on must fit within a descriptive framework rather than appealing to intuitive notions of the structure of discourse. There are essentially four levels of units, occurring in a hierarchical framework. The smallest unit is the act, then the move, then the exchange, and at the highest level the transaction. All these units are defined in terms of their function.

I now turn to some theories of the interaction of conversationalists to coordinate their respective control of the conversational floor.

2.3 Turn-Taking

According to conventional conversation theory, turn-taking is a process which consists of a switching of the roles of speaker and listener between participants in a conversation. The definition of a turn seems to approximate to that of an utterance, but as I noted in section 2.2 this unit is theoretically redundant within a hierarchy of discourse structure. Moreover, in this research, the turn unit has been abandoned as a unit of analysis both for theoretical and practical reasons (see Chapter 4 for a full account of this). I shall however retain the terms 'turn' and 'turn-taking' in this thesis because of their extensive use in the literature.

The organisation of the alternation must be accounted for by a system robust enough to allow for relatively precise timing between turns - speaker changes may be

achieved with little overlap or inter-speaker silence.[3] The system must also be able to cope with turn-taking in many different contexts. For example, the number of parties involved in a conversation may vary considerably, not just from one conversation to the next, but also within a single conversation. These parties may not have any specified order in which they enter the conversational floor, and when they do their contribution can vary enormously in length and complexity. Complexity does not necessarily imply chaos, and certainly turn-taking appears to adhere a set of potentially well-defined conversational rules.

Wilson & Zimmerman (1986) note three principle strategies which have been proposed as a model of turn-taking: stochastic models, signalling models, and sequential-production models.

2.3.1 Stochastic Models

The stochastic approach is essentially a Markov or other probabilistic process, where the conversation is made up of different states such as someone speaking, or silence, or both parties speaking. The model then describes the probabilities of transitions from one state to another (e.g. Jaffe & Feldstein, 1970). This approach could yield good output as far as reproducing actual turn-taking data is concerned. However, it contributes little towards an explanation of how turn-taking works, and will not be considered further in this thesis.

2.3.2 Sequential-Production Models (The Sacks et al. Model)

2.3.2.1 Turn Allocation

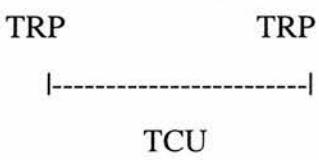
Sacks, Schegloff & Jefferson (1974) devised an ordered set of turn allocation rules which operate purely on a turn-by-turn basis to allocate a scarce resource (control of

---

[3]Figures vary from author to author, and from context to context, but one may say generally that any amount of overlap, or silence of more than approximately one second between speakers is less common than intervals of between 0ms and 1000ms.

the conversational floor). Each speaker is initially allocated a unit of control - *the turn constructional unit* (TCU). A TCU is proposed as the basic unit which interlocutors compete for in a conversation, and the boundaries of TCU's are called *transition relevance places* (TRPs):

5)



A TCU is roughly equivalent to a move, the difference being that whereas a move is defined in terms of the function it performs in the conversation, a TCU is only defined with reference to the TRP. A TRP may be identified using various criteria, and need not coincide with a move boundary.

A TRP is a point in a conversation where there is the *potential* for a speaker-switch. According to the rules of Sacks et al. the current speaker (C) may select a specific next speaker (N),[4] or may leave the floor open to any other speaker. If N does not self-select, then C may continue until the next TRP is reached, when again the turn transition rules come into play. The rules are set out explicitly below.

Rule 1 - applies initially at the first TRP of any turn

(a) If C selects N in current turn, then C must stop speaking, and N must speak next, transition occurring at the first TRP after N-selection

(b) If C does not select N, then any (other) party may self-select, first speaker gaining rights to the next turn

---

[4]For reasons of brevity, 'C' and 'N' will be used to refer to current-speaker and next-speaker respectively throughout the rest of this thesis.

(c) If C has not selected N, and no other party self-selects under option (b), then C may (but need not) continue (i.e. claim rights to a further turn-constructional unit)

Rule 2 - applies at all subsequent TRPs

When Rule 1(c) has been applied by C, then at the next TRP Rules 1 (a)-(c) apply, and recursively at the next TRP, until speaker change is effected

With these two main rules it is possible to account for several observations of turn-taking phenomena.

1) Inter-speaker intervals

An interval between the end of C's turn and the start of N's turn can be subcategorised into two types:

a) Interval between speakers

An interval between speakers occurs where C specifically selects N, and the turn passes to N at the first TRP after selection. One turn starts the moment the other finishes, so any intervals between speakers must fall inside N's turn. Rule 1(a) accounts for this situation.

b) Interval between turns

An interval between turns is any interval where C has not selected N, so if N does self-select and take a turn at speaking, that turn would not necessarily begin as soon as C reaches a TRP. Any interval between the two speakers would fall between two turns. Inter-turn intervals are accounted for by rule 1(b).

It is important to note here that in the rest of this thesis no distinction is drawn between inter-speaker intervals and inter-turn intervals, and the term *inter-speaker interval* is used as a cover-all term to describe any interval which falls between utterances by different speakers.

2) Within-turn intervals

Within-turn intervals occur when C reaches a TRP, but the floor is not claimed by another party so that C may continue with a new TCU. Rule 1(c) provides for these situations.

3) Lapses

If none of 1(a)-(c) are evoked, then a lapse in the conversation will follow. It is uncertain how much of an interval must occur to constitute a lapse, and clearly this will depend to a large extent on contextual and social factors.

4) Unintentional overlap

N may not calculate the location of a TRP correctly, and overlap unintentionally. 1(b) accounts for these situations, but only inasmuch as it predicts that N may choose to enter at a TRP even when not specifically selected to do so.

As well as selecting N, C can select the *type* of turn that N may use. This is achieved through use of another local management device - what Sacks (1972) calls an *adjacency pair*.

2.3.2.2 Adjacency Pairs

Any given conversation, so Sacks notes, is made up of at least two turns. Some of these turns are more closely linked to each other than to other turns, to the extent that the first turn determines the type of second turn which would be expected - for example, given a question one would normally expect an answer to follow. Sacks (1972) also notes that:

> a person who has asked a question has ... a reserved right to talk again, after the one to whom he has addressed the question speaks. *And* in using this reserved right he can ask a question.

Adjacency pairs form the basis of much conversation since they allow C to select both N and the next turn.

Sequences may also consist of different types of pair. Sacks notes that when C makes a Request, for example, there is the possibility of N making a rejection. Speakers attempt to reduce the likelihood of such a rejection by making what Sacks calls a *pre-sequence*. The pre-sequence determines whether or to what extent the request is going to be accepted. A typical pre-sequence would take the form:

6)

A: Are you doing much?
B: Well, not really. Why? ← Pre-sequence

A: Oh, well, I just wondered if you'd like to go out...

Often, the first part of a pre-sequence may be recognised for what it really is, and a response made which would normally be a reply to a notional main sequence. For example:

7)

1: Don't you think it's kinda cold in here? ← Pre-sequence
[2: Yes. Why?]

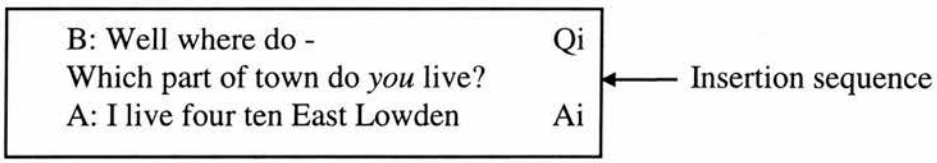
[3: Well, could you close the window?]
4: Sorry, I'm busy.

Turns 1 and 4 are the only ones that are uttered. The actual request would have occurred in 3, but is not necessary because the request was implied in 1. Turn 4 rejects that implied request.

Embedded sequences may also occur - what Schegloff (1972) calls *insertion sequences*. This is where N produces another first pair part rather than the expected second pair part. Its function is often that of a stalling device. Coulthard (1977) gives the following as an example:

8)

A: I don't know where the -wh- this address / / is. Q



B: Well you don't live very far from me A

Related to this is the *side sequence*, as proposed by Jefferson (1972). This is where the conversation is interrupted by an aside - the result of the need for clarification on some point. After the initial statement, a misapprehension follows, and then a clarification of the misapprehension. The conversation then picks up where it was broken. The main difference between a side sequence and an insertion sequence is that a side sequence does not begin with a first pair part, and the subsequent turns are not inserted, since there is no expectation of when they should end or of what should follow them. Another difference is that side sequences have an element after the initial misapprehension-clarification sequence, which operates as a 'full stop' on the side sequence by acknowledging that the misapprehension has been understood and the situation resolved.

Because a side sequence has no first pair part, the conversation cannot resume where it left off with a second pair part. Interlocutors therefore need a mechanism by which the conversation can be restarted. Jefferson notes two methods - *resumption* or *continuation*. Resumption represents the explicit demarcation of the return to the main conversation, and examples would be phrases such as 'Anyway' or 'So, where

was I?'. Continuation represents an implicit return, and typically attempts to hide the problem addressed by the side sequence. A continuation would take the form of words like 'so', 'and' or 'well'.

One might object to adjacency pairs on the grounds that first parts of pairs, such as questions, are not always followed by the matching second pair. But determination is not absolute, and N may decide not to follow with the expected part of the adjacency pair. If the rule or set of rules is broken, the result is socially marked in some way.

2.3.2.3 Appraisal of the Sacks et al. Model

A major drawback with the Sacks et al. model is its inability to account for timing in conversation. For example, rules 1(b) and 1(c) do not offer a fully explicit definition of concepts such as lapse or inter-turn interval, since no mention is made of the timing aspects involved, nor of the effect of change of conversational topic which may occur after a lapse, but not after an inter-turn interval. An account of timing needs to be specified because otherwise it is difficult to explain how, for example, C can become aware that N is not going to take the conversational floor, or how long C need wait to allow N to speak before continuing. Sacks points out that N can never *know for certain* when C has finished his turn, since it is always possible to continue or extend an utterance. N cannot therefore wait for that turn to finish entirely before beginning an utterance, and instead need only wait for points of *possible* completion.

Also, a distinction between overlap and intervals between speakers is impossible to make, and both cases are simultaneously accounted for by rule 1(b). The problem arises because the only difference between a positive interval and unintentional overlap results from the way a TRP is temporally projected or misprojected. Apart from not being able to distinguish the timing differences, the model is unable to account for how it is that N can project a TRP at all. It is just assumed that this is possible.

It might be possible to alter the turn allocation rules of the Sacks et al system. Rule 1(c) could be altered to include any case where N does not take up the chance to hold the conversational floor. Another rule could be inserted to account for use of the

back-channel, where C does not select N, but N nevertheless produces an utterance at a TRP. C would have to recognise this utterance for what it is - namely, *not* a claim for the conversational floor - and continue with the turn.

Rule 1 - applies initially at the first TRP of any turn

(a) If C selects N in current turn, then C must stop speaking, and N must speak next, transition occurring at the first TRP after N-selection

(b) If C does not select N, then any (other) party may self-select, first speaker gaining rights to the next turn

(c) If no other party takes the floor under options (a) or (b), or if another party makes use of the back-channel, then C may (but need not) continue (i.e. claim rights to a further turn-constructional unit)

Rule 2 - applies at all subsequent TRPs

When Rule 1(c) has been applied by C, then at the next TRP Rules 1 (a)-(c) apply, and recursively at the next TRP, until speaker change is effected

Another problem lies in distinguishing between intentional and unintentional overlap. Intentional overlap involves N deliberately not waiting for a TRP before self-selection. The reasons for this may be various, and need not be uncooperative. N might wish to take the conversational floor 'by force', without waiting for an appropriate and acceptable moment (a TRP) to take the floor. But as Tannen (1984) notes, intentional overlap may be used to indicate a sense of familiarity between participants in a conversation. Or it may be used by N to signal the level of understanding of C's utterance (Clark, 1996). It is in many ways a 'social glue' which can be used to form a bond between speakers. Quite how cooperative and uncooperative intentional overlap can be distinguished systematically is uncertain, but it may rely to some extent on the degree to which speakers are familiar with one

another, the topic of conversation, and paralinguistic factors such as loudness of voice.

The turn-taking rules of Sacks et al. have a further limitation in that they were based on observations from American English conversations.[5] Sacks, Schegloff & Jefferson (1974) suggest that there is an underlying rule which states that 'at least and not more than one party talks at a time.'[6] If too many people speak at once, all except one will typically yield quickly. If the opposite is true, and there is a silence (which one may refer to as an 'awkward' silence) one speaker will take the floor, either by actually speaking, or by some other linguistic gesture such as a filled pause, intake of breath, and so forth. It does not take too much reflection or experience of social situations to realise that while these observations may very often be true, conversation does not necessarily work in this way, even within one cultural framework, and let alone in different cultures. The notion that turn-taking is governed to a large extent by social factors is supported by evidence from some cross-cultural studies which have shown that cultures have different rules for who can speak when, rules for how many people may speak at once, and perhaps most noticeably, rules for how much inter-speaker interval (either positive or negative) is acceptable.

For example, Reisman (1974) notes that in Antigua someone may start speaking while someone else is already speaking, and the resulting overlap does not in itself signal a competition for the conversational floor. Albert (1972) reports that amongst the Burundi of Africa, turn-taking is determined by the social status of the speakers. But as Levinson (1983) points out, a similar process may occur in some situations in English-speaking cultures - for example in classrooms or meetings. More recent studies by Jennifer Coates, for example, indicate a highly complex structure to certain types of conversation within the same culture.[7] She has found that in conversation amongst groups of women, the normal rules of conversation and

---

[5]Although arguably all models of turn-taking suffer from the deficiency of focusing on one language and one culture.

[6]A somewhat curious way of saying that only one party may speak at a time.

[7]As outlined in a paper given at the Manchester Postgraduate Conference 1996.

turn-taking outlined by Sacks, Schegloff & Jefferson (1974) do not apply. Instead, overlap is common, and speakers often complete or repeat parts of one another's utterances. All this is often achieved without any sense of competition for a notional conversational floor. Such research has only been carried out relatively recently, and as such it is still uncertain to what extent this type of conversation is found in different cultures, and whether it is gender-specific. There is apparently some anecdotal evidence however that at least amongst groups of women non-competitive conversation occurs in cultures as diverse as European and Japanese. Men may, in some circumstances, use a non-competitive form of turn-taking. If nothing else, such studies should cause one to treat some previous studies with a degree of caution, and one is lead to wonder whether cultures which apparently have a greater tolerance of overlap in fact use forms of non-competitive conversation more often. It may be that all cultures have surprisingly similar allowances for the timing and coordination of turn-taking. What differs between cultures might be the circumstances that different styles of conversation are permitted in. That is, formal styles of conversation may require a certain rigidity of coordination not present in less formal styles, and this rigidity may be present across cultures. The apparent cultural differences may however arise because of the situations in which a formal style of conversation is permitted.

The Sacks et al. system of turn-taking rules provides a robust basis for accounting not only for specific cases of overlap, lapse, and silence, but also for the variability in conversation. Factors such as number of conversationalists, change in group size during the conversation, TCU size, or the lack of definition of turn length, can all be accounted for by the provision for local management under the rule system. Local management means that the system works purely on one transition at a time, regardless of what has come before or will follow. But there are many limitations to the system, centering on its inability to account fully for the projection of TRPs, and the timing of entries to the conversational floor, and its treatment of conversation as a rigid turnA-turnB system of exchange. The evidence from actual dialogue suggests that conversation is not that rigid, and that conversationalists accept a great deal of overlap, interruption and deliberate fade-out.

2.3.3 Signalling Models

The signalling approach is one which has arisen from a psychological tradition. It assumes that conversation is organised purely according to a series of cues which signal the closure of a turn. The emphasis here is on the word *purely*, because as I shall show later there certainly are cues which are used to project TRP location. The issue is whether these cues could be considered to act alone for an operative model of turn-taking, or whether cues could only operate as one aspect within a more comprehensive framework. According to the signalling approach, C signals an intention to hand over the floor (something which has been likened for the purposes of analogy to the 'over' announcement in some radio communication), and others may then 'bid' for a right to speak.

Duncan (1972) acts as a useful point of reference for the signalling approach. He notes six turn-yielding cues, and one or more of these cues may be displayed to give a turn-yielding signal. According to Duncan once the signal has been given and N acts to take the floor, then C should immediately yield the floor if the turn-taking mechanism is adhered to properly.

2.3.3.1 Six Turn-Yielding Cues

i. Intonation

Any pitch level terminal junction combination other than a flat intermediate pitch level junction acts as a turn-yielding cue. This would explain how in British and American English (and presumably other dialects and languages also) a phrase may be perceived as left 'hanging' when a speaker ends a turn but fails to use a rising or falling phrase-final intonation pattern. Given insufficient other cues to turn-yielding, listeners may fail to predict the end of the turn accurately, if at all.

ii. Paralinguistic drawl

A drawl on the final syllable (or stressed syllable) of an intonation-group. In other words, phrase final lengthening is said to act as a turn-yielding cue. Phrase final

lengthening is a well-accepted phenomenon, and is quite probably a language universal (Cruttenden, 1986).

iii. Body motion

The end of a gesticulation that is used during a turn may signal the end of that turn. Duncan (1972) defines a gesticulation as 'those hand movements generally away from the body, which commonly accompany, and which appear to bear a direct relationship to, speech' (p. 287) Excluded are those movements classed by Ekman & Friesen (1969) as self-adaptors and object adaptors. The former involve contact with self, such as brushing one's trousers, rubbing one's chin, and so forth. The latter involve movements concerned with objects that are being held or are in the immediate vicinity - for example a pens or paper. De Long (1974) shows a significant link between body movements and speaker change. He found that two movements - leftward motion of the head and downward motion of the head, arms or hands - would signal a turn closure when the movements occurred together or almost together. This does not mean that every such combination of movement acts as a direct signal to the next speaker. De Long claims that body movements reinforce linguistic cues to turn closure at TRPs.

iv. Sociocentric sequences

This term was originally used by Bernstein (1962), and refers to the use of certain stereotyped expressions, such as 'you know', 'you see' or 'like', which generally follow a statement and which add no further information to that statement. It remains uncertain, however, whether the use of sociocentric sequences actually aid the projection of a TRP, or whether they can be used to reinforce the existence of one once it has occurred, because it could be argued that the TRP falls before the sequence.

v. Paralinguistic pitch or loudness

A drop in pitch and or loudness in conjunction with a sociocentric sequence.

vi. Syntax

The completion, or near-completion, of a grammatical clause acts as an important turn-yielding cue.

2.3.3.2 Gaze

Further to the list of gestural cues listed above, gaze is an important factor in turn-taking. However, it is one which is difficult to measure or quantify. For example, Duncan & Fiske (1977) mention the problems of deciding on what counts as gaze, and whether eye contact is made. In particular the treatment of gaze across turn boundaries is problematic. Nevertheless, some analysis is possible (and one would argue necessary) for a model of turn-taking.

Kendon (1967) found that listeners generally look more at their conversational partners than speakers do, and will look at the speaker for relatively long periods, with relatively short glances away. Speakers, on the other hand, tend to look at listeners less, with the balance between gazing and looking away more equally balanced. Duncan & Fiske (1977) found that speakers gazed at listeners about 60% of the time, while listeners gazed at their partners about 87% of the time.

According to Kendon the close of a turn may be signalled to some extent by the speaker looking more at the listener - thereby signalling a willingness to switch to the role of listener. As this happens, the listener picks up on this and other signals (linguistic and gestural), and starts to gaze less at the speaker. In turn, the speaker, who is now gazing more at the listener than before, is able to detect the signals sent by the listener, and becomes aware that the listener has accepted the offer of the floor. Gaze signalling is therefore thought to be a highly interactive process near a TRP. Kendon also notes that fewer than one-third of the utterances which ended with gazing were followed by silence. Without gazing, nearly three-quarters of transitions had some form of delayed response.

Gaze (and gesture) do not in themselves mark the TRP so much as signal to the listener the speaker's *willingness* to yield the floor at any given TRP. It may well be as a result of this that Kendon found that in hesitant speech speakers spend far less

time gazing than during fluent speech. It should be noted that pauses and fillers (such as 'erm' or 'um') in hesitant speech are points where loss of floor are likely (Ferguson, 1975).

2.3.3.3 Appraisal of the Signalling Approach

No direct mention is made of whether and to what extent these cues are signalled before the TRP. But one might suspect that prosodic cues are likely to provide information about an imminent TRP. And if the work by Kendon is valid, then gaze patterns may be used to form smooth transitions between speakers by giving cues to a TRP before one occurs. It does seem clear that the interaction of different cues to TRP projection and indicators of TRP presence is highly complex, since it involves in itself complex issues such as gaze and gesture, pragmatics, and prosody. I shall not tackle this issue further here. Suffice it to say that these cues do exist, and that if a fuller understanding of turn-taking is to be gained then this is certainly one of the main issues which needs to be resolved. I now turn to the timing component in turn-taking, and to the matter of how, given the presence of the cues already outlined, it might be possible for N to time the start of the turn without undue pause or hesitation.

A further objection is that the signalling approach does not in itself provide a reliable account of the occurrence of overlap (both intentional and unintentional), lapse, and inter-turn and inter-speaker intervals, nor of how N may be selected - all of which are either accounted for or mentioned by the sequential production model. It provides a principled framework on which to base many observations, but does not attempt to make an account of conversation as a social process.

2.4 The 'Slot' Model

Wilson & Zimmerman (1986) claim that the sequential-production model is the only interactive model of turn-taking, and that the stochastic and signalling models involve the treatment of inter-speaker intervals as basically a simple response latency on the part of the next speaker. On this basis they separate the stochastic and the

signalling models from the sequential-production model. They adopt the Sacks et al. system as a basis for accounting for the timing of turn-taking.[8]

Wilson and Zimmerman's prediction is that with a stochastic or signalling model, the frequency of inter-speaker intervals would decline monotonically with duration, as shown in Figure 2. They do not make it clear exactly why such a pattern should be expected, although they do point out that Brady (1969) found this kind of relationship. It would seem likely that the intervals fit a skewed normal distribution.

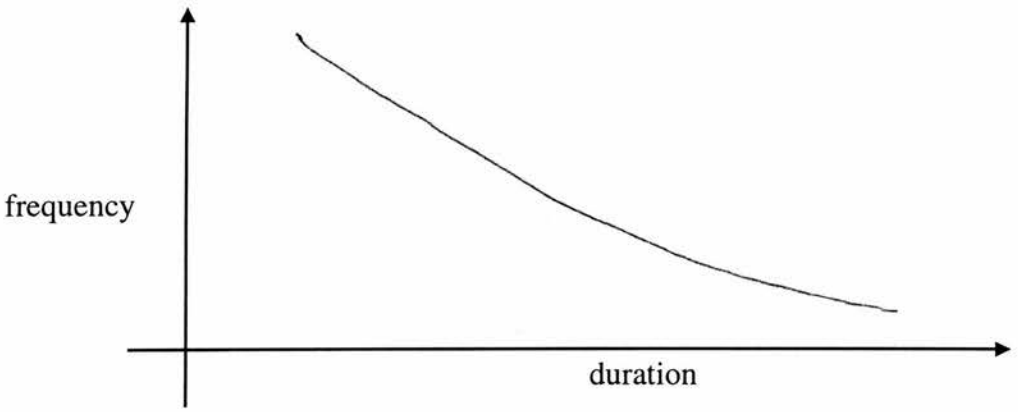


Figure 2 - Theoretical distribution of inter-speaker intervals, according to the stochastic and signalling models of turn-taking

However, for current consideration what is important is more the smoothness of the distribution. Wilson & Zimmerman claim that the sequential-production model differs from the other two in that it predicts a *periodic* pattern to the durations of intervals, as Figure 3 below shows. This prediction arises because the turn-taking rules of Sacks et al. are based in real time. The interval during which each conversationalist has the option to take the floor (i.e. exercise rules 1(b) and (c)) is termed the *slot.* Their hypothesis was that given that the cycling of options at any given TRP in a conversation occurs at a regular pace, then a fixed amount of time

---

[8]In fact, they used results from an analysis of inter-speaker intervals to provide evidence to support the Sacks et al. model, but their data could be used for a timing component in a general model of turn-taking.

will be needed for each cycle. It follows that if this were the case, one would expect to find a periodic pattern to the interval between two speakers, where each period equals the slot length. They argue that if one speaker reaches a TRP, and there is silence before N takes up the floor, then both rules 1(b) and (c) might have been passed up several times before rule 1(b) is finally exercised - that is, before N takes up the offer. The interval would then last 2Sk ms, where 2S is the slot length needed to pass up both rules 1(b) and 1(c) once, and k = 0, 1, 2, ... and is the number of times the two rules have been cycled through before N starts speaking. Hence the pattern shown in Figure 3, where the overall decline in frequency can be observed with an increase in interval duration, but only discrete interval durations are predicted.

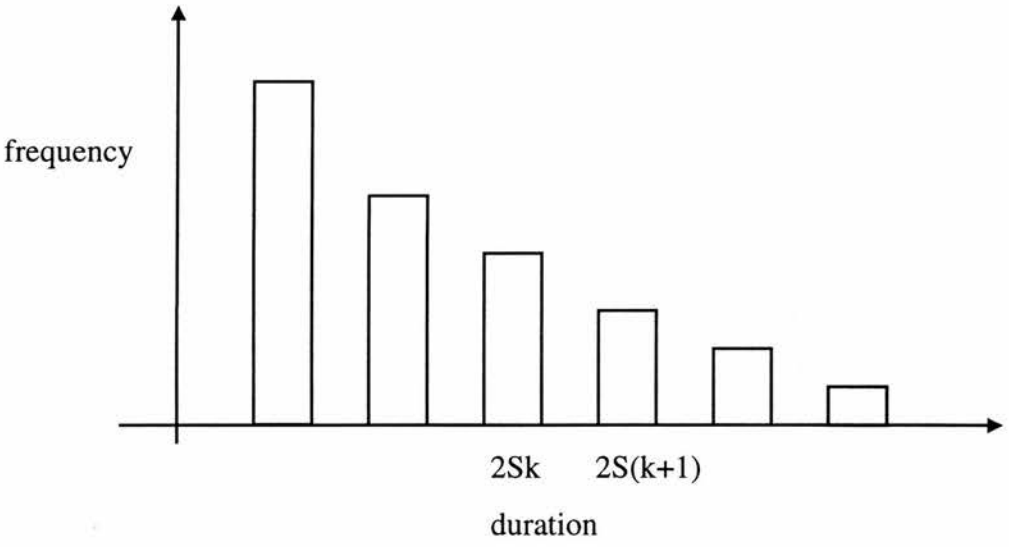


Figure 3 - Theoretical distribution of between-turn intervals according to the sequential-production model of turn-taking

Of course, this is a theoretical construct, and one would expect a pattern more like that in Figure 4 below.

Figure 4 - Predicted distribution of between-turn intervals

Wilson & Zimmerman analysed several hundred interval durations. These were taken from seven conversations, which themselves were taken from a corpus of conversations in a laboratory setting between randomly paired, unacquainted university students. Four of the seven conversations were male-female, two were female-female, and one was male-male. No subject took part in more than one conversation. A 9-minute segment was used from each conversation, and after analog-digital conversion were stored and processed on a PDP 11/45 computer. Each speaker's channel was processed separately. Silence was defined as a signal below a previously determined threshold noise level. Silence was ignored when the signal fell below the threshold for less than 100ms within one speaker's channel. Above-threshold noise lasting more than 100ms, and which was surrounded by silence, was converted to silence. Finally, the two speaker's channels were combined. A between-speaker silence was defined as a period of 10ms or more when both channels were below threshold. Periods of simultaneous speech were not used.

Wilson & Zimmerman found a cyclical pattern for between-turn interval durations, with a slot duration somewhere between 40 and 90ms, with a mean of 60 ms. As they point out, this would be 'implausibly short as response times if the mechanism involves one party waiting until it is clear that the other has not taken up the option to speak in his or her slot before initiating the speech process...' (p. 388).

That is, N would often be unable to start speaking quickly enough after only the first cycle of rules 1(b) and (c) has passed to be able to get the floor. So this is seen as further evidence of the projective nature of a hypothesised turn-taking mechanism. Wilson & Zimmerman propose that with a projective mechanism C is able to presume that even a slot length of as little as, say, 60 ms is long enough for N to take the floor if desired. They also suggest that if a prospective speaker were to start to plan to speak, then all that would be required if another speaker were to start before him or her would be to inhibit the incipient speech process. No relevant data seems to exist concerning the time required to shut down the vocal apparatus, but Wilson & Zimmerman suggest that this might be much shorter than the time required to initiate the speech production process from planning stage to the articulation stage. A possible means of testing this empirically might be to consider breath intakes in turn-taking. In fact, one could argue that all considerations of turn-taking should include an account of breath-intake, since it is this which signals intention to speak, and could be regarded as the actual start of a turn rather than the start of a speech signal. That this area has been largely disregarded in the literature thus far is almost certainly due to the practical difficulties in measuring breath intake immediately prior to the start of a speech signal. It is certainly an area worthy of future research.

One wonders whether a slot length may not simply fall within certain limits, and does not adhere strictly to a periodic effect. In fact, the theory relies heavily on the assumption that each cycle through the rule set will take an equal amount of time. But this need not be so. One could imagine that each successive cycle might take up less time than the one before - an assumption which seems quite plausible in view of the often competitive nature of turn-taking. On the other hand, each cycle might take more time than the one before, since each interlocutor may become more certain that the other is not about to speak, and that what was initially thought of as an interval is about to (or has) become a lapse. Clearly, the situation is complex, and it cannot be presumed as a basis of a model that each cycle must fit a certain duration to within a matter of only a few tens of milliseconds.

A similar argument concerning the variability of slot duration is mentioned by Wilson & Zimmerman, and is concerned with variations within a single conversation

caused by speech rate. That is, if the tempo of speech surrounding an interval is fast, then a listener may perceive the duration of that interval to increase. Whether true or not, there must be some doubt about the rigidity of a notional slot length within a conversation, and by extension across different conversations and different speakers.

There is also the argument that much conversation may not pass through the turn allocation rules more than once, since failure on the part of N to take the floor on the first pass would be sufficient either for C's continuing to speak, or for the realisation on the part of both interlocutors that a lapse has been reached, and that the topic of conversation may be changed. A lapse in a conversation therefore does not fit well into the Sacks et al. model. One would imagine that lapses are socially determined to a far greater extent (the 'awkward silence'), and that their duration varies considerably.

Nevertheless, whatever the actual slot duration, and whether there is a periodic nature to its occurrence, the distinction between rules 1(b) and (c) necessitates the establishment of some basic slot duration. The Sacks et al. model requires a definition of how long an interval may be tolerated following any given TRP, before N's potential opportunity to take the floor passes and C continues. Of course C need not continue, and the floor may then become open to N again. There is also a need for an account of how long C would wait under rule 1(a) for N to start speaking, before C would resume or at least attempt to reselect N.

The slot model will not be considered further in this thesis. It is felt that while it is an interesting model, and makes a contribution towards integrating social and cognitive factors in conversation, it suffers from making assumptions about the temporal rigidity of each of the cycles of the Sacks et al. turn allocation model.

In the next section there follows a discussion of the *rhythmic coordination hypothesis* (e.g. Couper-Kuhlen, 1993). This, like the Slot Model, is founded on the basic sequential-production view of turn-taking. Its unique feature is that it presumes that the coordination of turn-taking is achieved through the perception of rhythmic sequences in utterances.

2.5 The Rhythmic Coordination Hypothesis

The rhythmic coordination hypothesis (Couper-Kuhlen, 1993) claims that:

a) speech may be said to consist of rhythmically arranged perceptual beats;
b) N is able to pick out the rhythmic beats present in C's utterance;
c) once N has located a TRP, N's entry point is by default timed to coincide with the next beat after that TRP.
d) if the entry point does not coincide with the beat then this may be used as a signal of some social or cognitive difficulty.

The advantage of this approach (and also of the slot model) is that it adopts a social-interactive view of conversation, combined with a 'cognitive' element. That is, it attempts to combine an account of the social relationships between a group of individuals with an account of how those individuals carry out the finer detail of the coordination process itself (through the use of rhythm).

2.5.1 Perceptual Isochrony

Isochrony can broadly be described as the separation of several of the same type of events by equal periods of time. Rhythmic beats in music can therefore be regarded as isochronous. The notion of isochrony in *speech* first appeared in a modern linguistic sense in the early twentieth century, when Daniel Jones (1956) noted that the length of a vowel in a stressed syllable reduced if it was followed by unstressed syllables. He suggested that one of the possible causes for this was the observed tendency in English for stressed syllables to follow one another at more or less equal temporal intervals. This idea of (relatively) strict productive isochrony was maintained, indeed extended, by later researchers (Pike, 1945; Abercrombie, 1967; Halliday, 1985). More precise acoustic analysis clearly shows that stressed syllables in English are not spaced at regular intervals (Classe, 1939; Lehiste, 1977).

However, acoustic measurements of isochrony do not take account of the *perception* of isochrony. Lehiste (1977), Donovan & Darwin (1979) and Darwin &

Donovan (1980) found evidence that listeners tend to impose a regularity on an utterance, even where there is no strict acoustic regularity. Isochrony may therefore be a *perceptual* phenomenon, based on the acoustic signals of perceptually prominent syllables. While these are not isochronous in any strict sense, they may have a certain degree of periodicity sufficient for perceptual isochrony.

Further to this, evidence from Scott, Isard & de Boysson-Bardies (1985), Roach (1982), and Dauer (1983), seems to suggest that the traditional dichotomy of stress-timed versus syllable-timed languages, as maintained by Pike (1945) and Abercrombie (1967), is no longer valid. To assume that isochrony can only be present in stress-timed languages such as English, German or Russian, is therefore incorrect. The conclusion drawn by Dauer (1987)[9] is that the tendency for stresses to appear to occur regularly in an utterance is a *language-universal*, and that rhythm should thus be seen as acting along a continuum, where the differences in perceived rhythm are caused largely by phonological factors (the syllable structure, or reduction of vowels in unstressed syllables, for example). If this is the case, the perception of isochrony in speech would be universal.

An understanding of speech rhythm requires an understanding of the relationship between perceptual isochrony and actual acoustic data. Huggins (1972a) found that subjects were sensitive to changes in segment duration, and that these changes had affected perceived stress and perceived rhythm. Importantly, when subjects based their judgements on changes in perceived stress or rhythm they were often able to detect smaller changes in duration than when they attended to other aspects. In a further study, Huggins (1972b) tested whether a change in the duration of one segment should be compensated for in an adjacent segment if the sentence is to remain fluent. The results indicated that compensation occurs between some segments, but not others. This depended on where the two segments occur with respect to word and syllable boundaries. Thus, Huggins concluded that the perception of timing in speech is based on events at the syllable level rather than at the

---

[9] Cited in Couper-Kuhlen (1993).

segmental level, and that it is important to maintain the rhythm of the sentence if the sentence is to sound temporally fluent.

2.5.2 Rhythm and Cognitive Processing

The relationship between acoustic data and the (often strong) subjective experience of perceived isochrony is uncertain therefore. A possible cause, Couper-Kuhlen (1993) suggests, may be the presence of *rhythmic gestalts*. The phenomenon of the gestalt (or pattern-recognition, as it is now conventionally called in the cognitive sciences) is common in perception. Behind this principle lies the need for the perceptual system to segment input - to make sense of a mass of data by 'chunking' it into discrete packages.

Couper-Kuhlen (1993) points out that speech may be said to be articulated in perceptually discrete packages, where some of these packages are perceptually more salient than others. Each salient syllable and the non-salient syllables around it form a group, where each group can combine with other groups to form higher-level groups in a hierarchical structure. An isochronous sequence may therefore be thought of as a series of low-level groups which combine together into a perceptual whole at a higher level. Couper-Kuhlen claims that the principles used in organising isochronous sequences are based on a set of principles, which need not all be present for isochrony to be perceived, but which nevertheless produce the strongest perception of isochrony when they are all present. They are:

a) *Proximity.* The closer that salient syllables are in time to one another, the more likely it is that the sequence of salient syllables will be perceived as a whole.

b) *Similarity.* According to this principle, syllables with the same type of prominence form salient sequences. A series of pitch accented syllables would therefore form a 'natural' group. The number and type of less prominent or unstressed syllables should also be a factor, such that a sequence which contains a similar number of unstressed syllables between each stressed syllable would be likely to be perceived as an isochronous sequence.

c) *Objective set.* A sequence may be perceived as isochronous according to context. A sequence of otherwise non-isochronous prominent syllables may be perceived as isochronous if there is a sequence of isochronous prominences preceding it.

d) *Direction.* Even if successive intervals gradually shift in tempo, they may nevertheless still be perceived as an isochronous sequence.

e) *Good continuation.* It is to be expected that in a perceptually isochronous sequence the location of the next prominent syllable is entirely predictable from the rhythmic pattern of preceding prominent syllables in that sequence.

f) *Closure.* An isochronous sequence requires at least three prominent syllables before the rhythmic structure can be considered complete.

The treatment of perceptual isochrony as a form of pattern recognition, or rhythmic gestalt, attempts to explain why prominent syllables are perceived as falling within groups, and are perceived to be equally separated in time within those groups. However, there are three problems with the account.

First, it is not clear from a gestalt account why a sequence of prominent syllables should be perceived as being equally spaced in time. The proximity principle simply states that the nearer prominent syllables are to one another, the stronger the sense of isochrony.

Second, it is not clear how often, or whether, perceptual isochrony actually occurs in speech. Couper-Kuhlen (1993) claimed that a small group of expert subjects could separate isochronous from non-isochronous sequences in speech through multiple listenings (see section 2.5.6), although there is some reason to question this because of the (necessarily) subjective nature of the decisions involved, and the relative difficulty with which they were reached (everyday conversation obviously does not allow the luxury of multiple listenings and time for deliberation over whether a sequence is isochronous or not).

Third, this approach seems to presume that a sequence of prominent syllables is bound together by its perceived *rhythmic* properties. But other factors are involved in grouping linguistic sequences together. Syntactic, semantic, and intonational factors, amongst others, act to bring perceptual structure to a sequence of speech.

The tendency for the mind to group non-discrete signals into packages may be accounted for by a gestalt approach to perception, where segmenting signals into discrete units may facilitate perceptual processes. In turn, it is claimed that this may explain why it is possible for speech to be broken into rhythmic units, although it is not clear why the grouping of speech into rhythmic units should be any more beneficial than the grouping of speech into other perceptual units. Furthermore, there is no evidence, even if speech is frequently perceived as isochronous, that rhythm does in some way aid perceptual processes. But it is not certain whether perceptual isochrony is at all common in conversation, or whether rhythmic units are particularly salient. Therefore, the drawback with the rhythmic coordination hypothesis is that it takes for granted that rhythm and isochrony are major factors in conversation, when there is no direct evidence that this is the case.

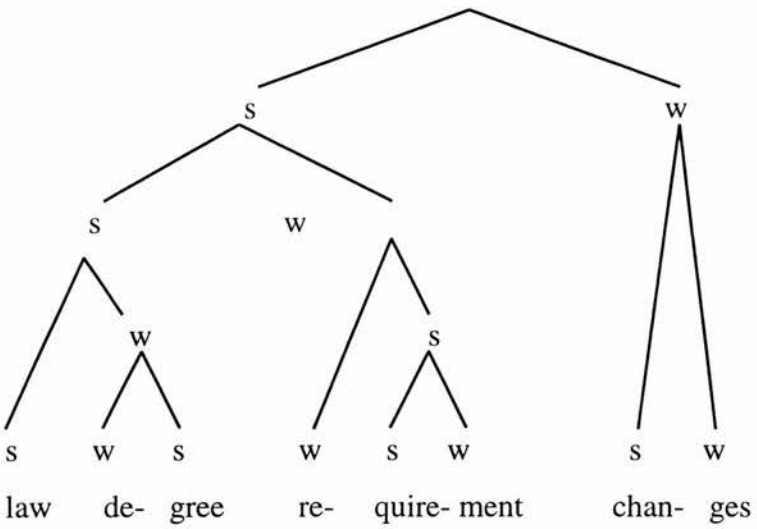
2.5.3 The Hierarchical Organization of Speech Rhythm

One of the main problems with a rhythmic approach to the timing of turn-taking is to formulate a set of principles that can be used to determine which syllables in an utterance are stressed. It is well known that stress is primarily a perceptual phenomenon (e.g. Cruttenden, 1986), and that as yet there is no full understanding of the way that various acoustic properties of speech, such as pitch change, loudness, duration, interact to produce the perception of stress. Moreover, semantic and syntactic considerations must play a role in the perception of the location and extent of stress. In essence then, the problem is to generate a consistent view of rhythm.

Couper-Kuhlen (1993) used a metrical, or hierarchical, approach to rhythm (e.g. see Liberman & Prince, 1977; Selkirk, 1984; Hayes, 1984; Giegerich, 1985). Although several approaches have been adopted or proposed in the last twenty years, for the purposes of explaining the basic concepts I mention only the original work on

this area (Liberman, 1975; Liberman & Prince, 1977). Here, the hierarchy of relative prominences was expressed in the form of *metrical trees.*

9)



(Liberman & Prince, 1977)

The trees relied on a binary relation between *strong* (s) and *weak* (w) syllables and groups of syllables. This binary relation was also expressed at a morpho-syntactic level. The metrical trees were then mapped onto *metrical grids.* These are matrices within which each column represents one syllable in an utterance. Each column consists of at least one grid mark (usually represented with *x*). More grid marks are added in such a way that the strongest syllable within every *s* constituent in the metrical tree must have one more mark than the strongest syllable in its sister *w* constituent. The relative strength of each syllable is represented by the height of the column. The grid corresponding to the tree in example 9) is:

10)

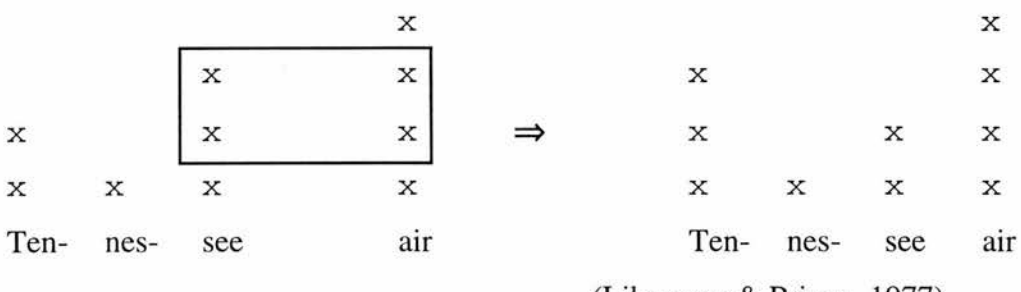
```
x
x           x           x           x
x     x     x     x     x     x     x     x
law   de-   gree  re-   quire-ment- chan-  ges
```

(Liberman & Prince, 1977)

The hierarchical approach to stress can help to explain certain phenomena in speech. The most notable of these is *stress shift.* This is the phenomenon which can sometimes occur when two stresses occur next to, or near, one another. The stresses appear to be shifted so that they are further apart, and no longer 'clash'. For example, the normal stress placement for the word *thirteen*, uttered in isolation, would be on the second syllable. Yet it is quite common for the stress to shift to the first syllable when *thirteen* is followed by the word *men.* A simple linear model may be able to account for some of these examples of stress shift, but in many cases stress shift has been observed when two stressed syllables were separated by several unstressed syllables.

The solution to this from a metrical perspective is to state that for stress to move, a clash must occur between two syllables at the same level on the grid such that there are no intervening syllables one level lower (with an additional rule governing the number of possible intervening syllables at lower levels), as shown in 11) below.

11)



(Liberman & Prince, 1977)

In 11), there is a clash between the syllables '-see' and 'air'. This is resolved by the movement of a beat from '-see' to 'Ten-'.

The number of levels in the metrical grid vary from one theorist to another. However, a common form of prosodic hierarchy is that found in Nespor & Vogel (1986, 1989), and it is this that was adopted by Couper-Kuhlen as a basis for the rhythmic coordination hypothesis. Here the hierarchy consists of the syllable (SYL), the foot (FT), the phonological phrase (PHR), the intonational phrase (I) , and the phonological utterance (U). Each of these units consists of one or more units at the level below it in the hierarchy, where the syllable level is the lowest. Couper-Kuhlen argues that English speakers may achieve isochrony at different levels of a grid. Certainly, it is not clear from the literature quite how a timing element is to be incorporated into a metrical model, and therefore how isochrony could be accounted for. As Couper-Kuhlen points out, there is a presumption in the literature (e.g. Selkirk, 1984) that the beats in a metrical grid represent constant time-values. Yet there is a contradictory assumption that the lowest level beats ('demibeats') in a stress-timed language have variable durations. But if a timing component were included in the model, and beats on a metrical grid were not considered to occur at regular intervals, or if the model were treated purely as a theoretical construct, there remains the problem of accounting for isochrony. Couper-Kuhlen addresses this problem by assuming that isochrony can be set up at not just one level (say, the foot level), but at one of several different possible levels. While there may often be no isochrony at the basic beat level used by Selkirk, there may be isochrony at a different level.

One might wonder why it is that prominences are set up at certain levels and not others. Couper-Kuhlen's solution to this is the interaction of *tempo* with rhythmic structure. For a speaker to form a rhythmic pattern (to *rhythmize* according to Couper-Kuhlen's terminology) at a certain level would in turn require that a given tempo be set up. And yet this tempo might not be desirable - perhaps being too fast for the given utterance and context. In other words, there is a trade-off between the desired tempo and the level of rhythmization in the prosodic hierarchy, where tempo

takes priority. It is therefore possible for lower level (e.g. syllable level) beats to have variable realisation in time, but for a higher level to be arranged isochronously. For example:

12)

U	X					X				
I	X					X				
PHR	X				X	X				X
FT	X		X		X	X	X	X		X
SYL	X	X	X	X	X	X	X	X	X	X
	wel-	-come	mi-	-ssis	giles.	he-	-llo	mis-	-ter	hodge

(Couper-Kuhlen, 1993)

Here, Couper-Kuhlen claims that the stresses occur perceptually isochronously on *wel-*, *Giles*, *he-*, and *Hodge*. That is, perceptual isochrony is established in this instance at the PHR level. One may therefore deduce from this that beats at the lower levels of FT and SYL need not be temporally arranged, other than to 'fit' within the temporal structure established at the PHR level. However, see 13).

13)

U			X			X				X		
I			X			X				X		
PHR	X		X			X		X		X		X
FT	X		X	X		X		X		X		X
SYL	X	X	X	X	X	X	X	X	X	X	X	X
	how	d'you	do	ma-	-dam.	don't	be	for-	-mal.	Dick's	the	name

(Couper-Kuhlen, 1993)

We can see that isochrony has been established at either the U or I level. Couper-Kuhlen claims that it is the lowest level of two or more possible levels which is the relevant one for interlocutors - so here, the I level would be the pertinent one.

The rhythmic hypothesis makes use of a prosodic hierarchy, arranged on several levels. Perceived isochrony can occur at any of these levels, depending on the tempo of an utterance. However, a consideration of isochrony cannot be complete without taking into account the possible points in a syllable around which isochrony is perceived: so-called perceptual centres.

2.5.4 Perceptual Centres

In the last two decades there has been a line of research into what are called *perceptual centres* (or P-centres). The research stemmed from a study by Morton, Marcus and Frankish (1976) in which listeners attended to lists of monosyllables, and were asked to arrange the timing of items in the list such that they appeared to occur rhythmically. The acoustic waveforms of the (perceptually) rhythmically arranged monosyllables were then analysed. It was found that neither the syllable onsets nor the vowel onsets were aligned at equal intervals. When they were, subjects would report that the list did not sound rhythmic. Some other point in the acoustic signal was used as an 'anchor-point' by the listeners in their judgements of what appeared to them to be rhythmic. This anchor-point is the P-centre.

It should be noted that P-centre location appears to be systematic, and is not merely a product of the variation of duration permitted in acoustic signals in order for there to be a perception of rhythm. Morton, Marcus & Frankish (1976) were unable to determine precisely the factors involved in P-centre location, although they were able to find some correlation between P-centre location and the duration of acoustic energy before vowel onset. That is, where there was a long consonant at the beginning of a word, the P-centre was found to be located earlier in the syllable than when the consonant was short. Marcus (1981) also showed that P-centre location could be linked to vowel length and to the relative duration of the final consonant in a CVC syllable. The longer the vowel and the later the final consonant, the later the P-centre in the syllable. There are therefore two forces acting in opposition. The

duration of the prevocalic segment 'pulls' the P-centre toward the onset of the syllable. The duration of the syllable rhyme acts to pull the P-centre toward the offset of the syllable. Figure 5 shows the relative alignment of P-centers in a series of spoken digits.

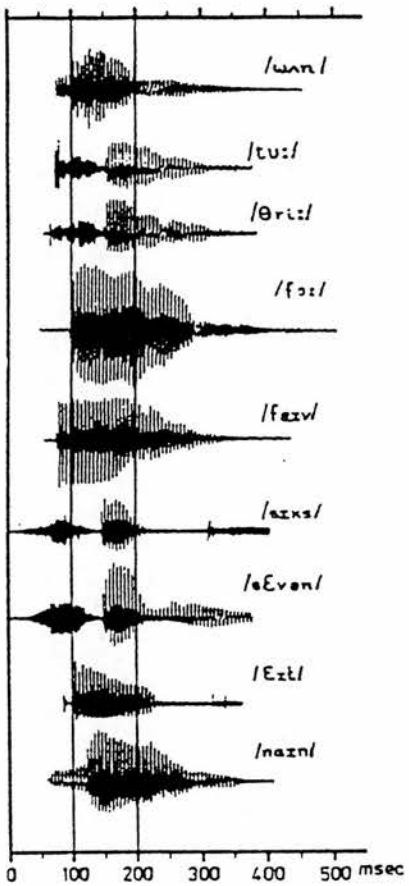


Figure 5 - The relative P-center alignment in the spoken digits '*one*' to '*nine*'. (Morton, Markus & Frankish, 1976)

Marcus (1981) suggested an algorithm for determining P-centre location in a CVC syllable, based on these two competing forces:

14)

$$P = \alpha x + \beta y + k$$

where P is P-centre location, x is the duration of the initial consonant, y is the duration of the syllable rhyme, α and β are parameters of the model, and k is an arbitrary constant which accounts for the fact that P-centres can only be determined for a stimulus *relative* to another.

The data from Marcus (1981) allowed values of α and β to be set at 0.65 and 0.25 respectively. However, the algorithm therefore becomes descriptive rather than predictive, and the values of α and β may well vary in different circumstances. Worse, k appears to be quite arbitrary.

Therefore, despite a considerable amount of research into P-centres (Fowler, 1979; Marcus, 1981; Cooper, Whalen and Fowler, 1986, 1988; Fox and Lehiste, 1987; Hoequist, 1983) no simple explanation of this phenomenon has arisen, and the most that can usefully be said is that P-centres result from a complex combination of the duration of the prevocalic segment, and the duration of the syllable rhyme. No predictive model has been produced. Until a working model of P-centre location is made, measurements of perceptual isochrony have to be based around the traditional methods of using either syllable or vowel onset.

2.5.5 Speech Rhythm as a Coordinator of Turn-Taking

Using the notion of rhythm acting at more than one level of the prosodic hierarchy, Couper-Kuhlen proposes that the rhythm established not merely as operating within one utterance for one speaker, but across turn transitions. The claim, it should be stressed, is not that this process always occurs, but that it can, and often does. There are therefore two cases that may be described - the unmarked case and the marked case.

2.5.5.1 The Unmarked Case

Here, the next speaker would time their entry to maintain the rhythm (and hence tempo) of the current speaker's utterance. For the next speaker to join, the current speaker would have to be speaking in a way that the listener could perceive as isochronous.

15)

```
                    TRP
A:    *....*.....*....
      ^     ^      ^     ^          ^
          I1     I2    I3    I4

B:                        ....*    .....*
```

Where * represents a syllable with prosodic prominence, ^ is used to indicate the pattern of beats set up by the first speaker, and . represents a syllable with no prominence. I1, I2 and I3 represent interstress intervals which are *perceptually* isochronous. The important point here is to note that I3 is perceptually equal to I1 and I2 - the timing of turn-taking is governed by the rhythm of the current speaker's utterance. I4 need not be equal to I1 and I2 according to the hypothesis. It is also important to note that the number of unstressed syllables within I1, I2, and I3 may vary, and that some of I3 may consist of silence in an inverse relationship to the number of unstressed. Presumably there must be an upper limit to the number of unstressed syllables that may be 'squeezed' between prominences before the illusion of isochrony breaks down.

2.5.5.2 The Marked Cases

a) Anticipated and early onsets

16)

```
                    TRP
A:    *.....*.....*..
      ^     ^     ^     ^     ^
        I1    I2    I3    I4

B:                        ...*...*.....*....*
```

This is where the next speaker does not enter on the next beat after the TRP. Note that as before, rhythmic disintegration may ensue - B may enter too early, but compensates by coordinating successive prominences with the rhythmic beat set up by the first speaker (as in the diagram above). Alternatively, B may set up a rhythm which does not coincide with A's beat:

17)

```
                    TRP
A:    *.....*.....*..
      ^     ^     ^     ^     ^
        I1    I2    I3    I4

B:                        ...*.....*.....*
```

b) Delayed onsets

18)

```
                                    TRP
A:      *....*......*..
        ^       ^       ^       ^       ^
            I1      I2      I3      I4

B:                                                  ..*.....*
```

Here, rhythmic coordination is said to be maintained, although one beat is 'missed' by B.

c) Late onsets

19)

```
A:      *....*......*..
        ^       ^       ^       ^       ^
            I1      I2      I3      I4

B:                                  ..*.....*
```

This situation is the same as b), only speaker B's late onset is not coordinated with the rhythm set up by A.

2.5.6 Empirical Evidence

The hypothesis, then, is that isochrony is perceived in speech both within and across speaker turns. To test this, Couper-Kuhlen took a two-minute fragment from a radio phone-in conversation. The first step was to determine which syllables were stressed at the utterance level. This immediately poses a problem, because stress is primarily a perceptual phenomenon. Although it is known that the perception of stress depends on vowel length, pitch, and amplitude, it is not known exactly how these factors combine. This makes the reliable instrumental separation of prominent syllables from non-prominent syllables difficult. Couper-Kuhlen used two trained native speakers to mark prominent syllables auditorily.

The next step was to determine which of the prominences formed perceptually isochronous sequences. Again, this had to be assessed auditorily. Couper-Kuhlen used two trained native speakers to make the decisions. Foot or finger tapping was used as a strategy to facilitate the assessment.

This analysis revealed that the conversation consisted of both isochronous and non-isochronous sequences. The isochronous sequences were found to extend both across intonation boundaries within a speaker's turn and across speaker switches. The conclusion from this was that the perception of isochronous sequences cannot be explained purely in terms of intonational, syntactic, or semantic considerations, and that it must operate at a different level.

The data was therefore split into isochronous and non-isochronous sequences. Couper-Kuhlen then carried out an instrumental analysis of the data, to determine a cut-off point for the variation in inter-stress duration that could be tolerated before an isochronous sequence would be regarded as an anisochronous sequence. She found that for isochronous sequences inter-stress intervals ranged from 0.21 sec. to 1.2 sec. For non-isochronous sequences, the inter-stress interval ranged from 0.1 sec. to 1.9 sec. Generally, less variation in inter-stress interval is allowed for the perception of isochronous sequences than for the perception of non-isochronous sequences.

But of real importance and interest to the estimation of permissible deviation of inter-stress intervals in an isochronous sequence is an analysis of *relative* durations. So, in an isochronous sequence the question arises whether an inter-stress interval is similar in duration to surrounding intervals. There are two ways of estimating variation. One approach is to compare each inter-stress interval with the mean inter-stress interval of the preceding intervals in that sequence, and calculate the percentage difference between the two. The other approach is simply to compare each inter-stress interval with the immediately preceding interval. Couper-Kuhlen used the latter method, claiming that it had the advantage of not treating intervals as 'immutable parts of a chain' (p. 57). This is problematic, however, because presumably to treat intervals as part of a chain is what makes those intervals appear to be isochronous.

Couper-Kuhlen found that a comparison of percentage differences for isochronous and non-isochronous sequences revealed that no isochronous interval differed from a prior interval by more than 40%. Her results are shown below in Figure 6.

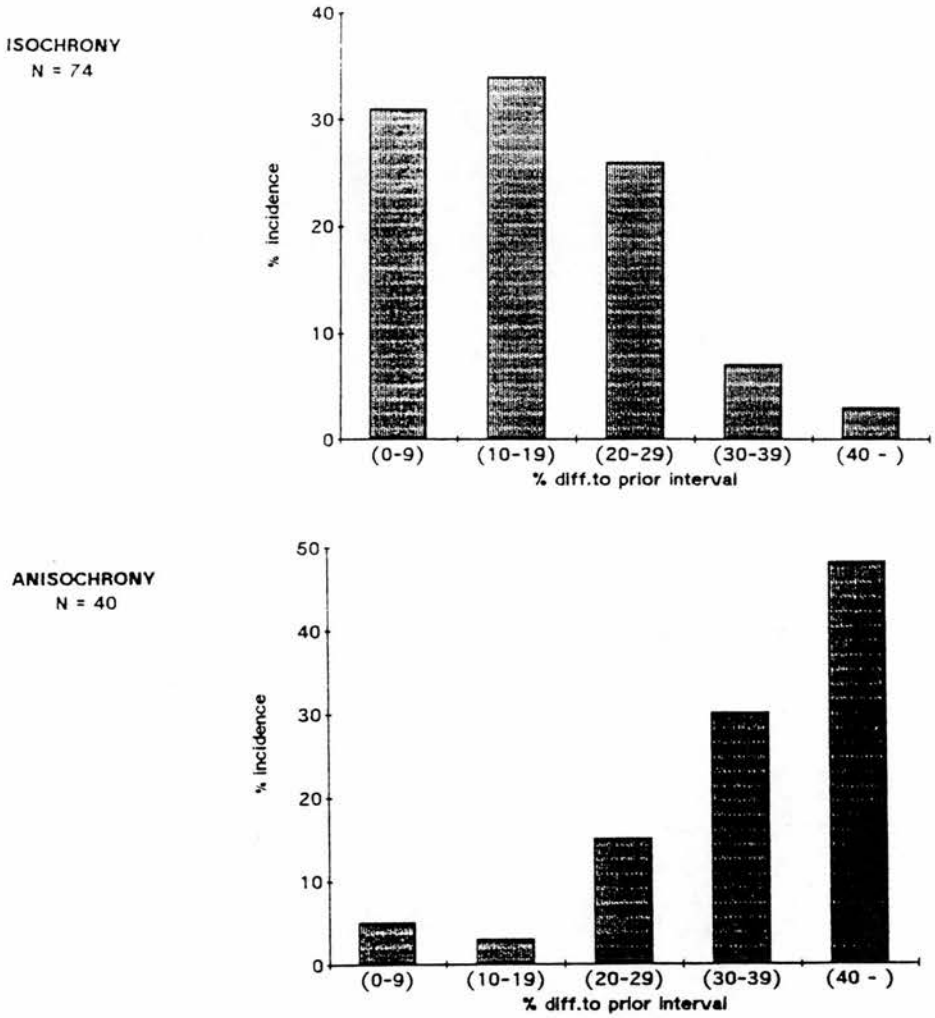


Figure 6 - Percent differences in interval duration with respect to a prior interval for isochronous and non-isochronous sequences. (Couper-Kuhlen, 1993: 58)

But almost a third of isochronous intervals differed by more than 20%. Interestingly, Couper-Kuhlen notes that in these cases there is often an intonation boundary present, and that over 50% of cases where there is an intonation boundary, the interval is longer than an immediately prior one. Also, it was found that many of the intervals which had an increase of 30% or more over a prior interval contained a speaker switch as well as an intonation boundary. Consequently, inter-stress intervals

which occur across speaker switches may generally be longer than other inter-stress intervals, and yet still fit within the limits of perceptual isochrony.

There are, however, some difficulties with this analysis. As Couper-Kuhlen admits, there are some intervals which are perceived as isochronous, and yet which differ by more than 20-30% from a prior interval. In these cases, Couper-Kuhlen claimed that the problem may be to do with tempo. That is, when intervals are very short, the permissible percentage difference from a prior interval may be greater than when the intervals are longer.

Another problem is that some intervals were perceived as non-isochronous even though the percentage variation in the inter-stress interval duration was less than 20-30%.A possible explanation for this may again be the presence of speaker switches. That is, pauses within an utterance may break down the perception of isochrony, even where interval variation is less than 30%. But a similar pause across a speaker switch may be tolerated, allowing the perception of isochrony.

2.5.7 Tolerance Levels for Perceptual Isochrony

The question arises that if perceptual isochrony exists, then how rigid does it have to be before it breaks down. Presumably the fragility of the perception of isochrony varies according to the mean intervals between stresses, although hard evidence is scarce both for the influence of tempo on isochrony and for the degree of freedom permitted for perceptual isochrony to occur.

Couper-Kuhlen (1993) carried out a pilot experiment to test the size of a possible rhythmic 'window'. A series of tokens was made, consisting of the word *sat*, repeated several times and spaced at regular intervals. Three sets were made, where each set had a different interval between tokens. The intervals between tokens were altered systematically by increasing the interval between the first and second token in the series, and reducing the interval between the second and third token, and so on. This produced four non-isochronous series, and one isochronous series, in each set. These series were presented to five subjects, who were asked to rate each series as either 'regular', 'irregular', or 'unsure'. Results from this suggest:

a) that there is a gradual shift between perceptual isochrony and non-isochrony. It does not therefore appear to be subject to the same sort of categorial perception that, for example, vowel onset times are.

b) that approximately a 20% shift from actual isochrony is required before a sequence of tokens are perceived to be either definitely or possibly non-isochronous.

c) that the permissible percentage difference between tokens before perceptual isochrony breaks down is the same at different tempos. Therefore, the amount of tolerance is affected by tempo, although this amount varies in proportion to changes in tempo.

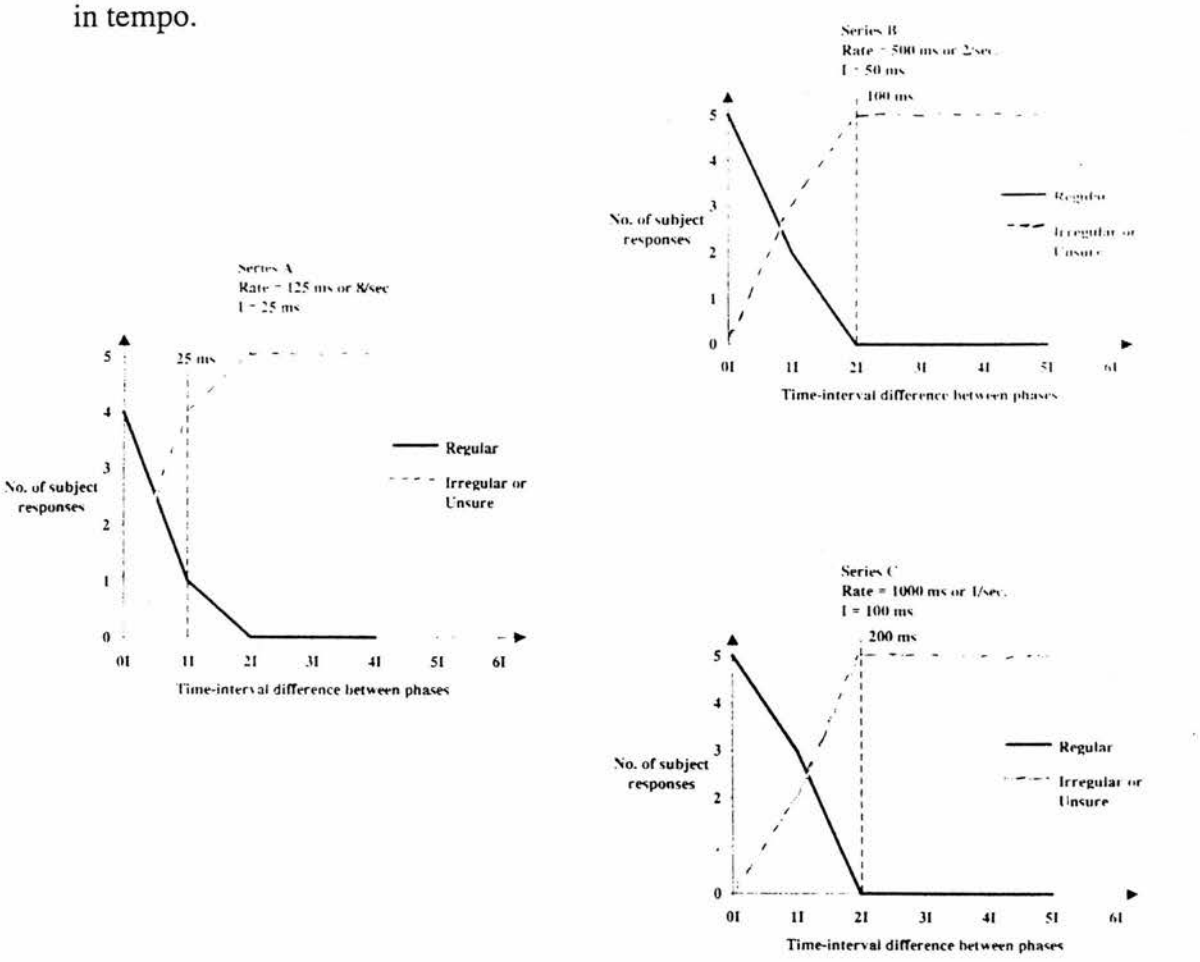


Figure 7 - Regular vs. Unsure/ Irregular judgements: tolerance zones for clear isochrony

2.5.8 Some Theoretical Advantages of a Rhythm Based View of Timing

2.5.8.1 Rhythm in Interaction

Couper-Kuhlen claims to have found evidence that not only is it possible for listeners to be able to judge utterances as being either isochronous or non-isochronous, but that the majority of turn transitions are perceived as isochronous. We have already encountered the conjecture that isochrony is perceived as a by-product of chunking signals into smaller units. But there is still the question of why any speaker should try to establish rhythmic coordination at turn transitions. There are two possible answers to this.

The null hypothesis is that rhythmic coordination occurs by chance. But because isochronous sequences were much more common than non-isochronous sequences, Couper-Kuhlen suggests that this hypothesis is unlikely to be true.

An alternative hypothesis is that there are temporal accommodation factors involved in conversation. Matarazzo et al. (1963) and Jaffe & Feldstein (1970) reported some congruence of mean inter-speaker interval between speakers in a conversation, where the mean interval varied when the same speaker conversed with different partners. Rhythm is one means by which accommodation can be achieved (Webb (1972) found that accommodation can be found in the speech rate of participants in a conversation). Under this view, the presence of isochrony across turn transitions functions to form a sense of mutual interpersonal influence in conversation.

A second alternative hypothesis is that rhythm may act as *a contextualization cue*. A contextualization cue signals the context in which an utterance is to be interpreted. Erickson & Shultz (1982) claimed to have found a rhythmic organization in verbal and gestural interaction, and moreover that the rhythmic structure of speech and gesture acts as a contextualization cue. In other words, rhythm can signal the social relationship, or shared interpretive framework, of participants in a conversation. As Erickson & Shultz (1982) note:

> What is suggested is that behavioral regularity, especially rhythmic regularity, may be prima facie evidence of shared interpretive

> frameworks among those engaged in interaction. The absence of such rhythmic regularity may be evidence that there is inadequate sharing of interpretive frameworks. (p. 143)

They identified four types of rhythmic instability:

a) Individual rhythmic instability - when an individual's speech is not rhythmic.
b) Mutual rhythmic instability - when the periodicity of the regular rhythmic interval is disturbed by the behaviour of both participants.
c) Mutual rhythmic interference - when the speech of each conversationalist is rhythmic, but there is no coordination of their rhythms.
d) Mutual rhythmic opposition - when the rhythmic coordination in the behaviour of the participants disintegrates.

2.5.8.2 Speech Rhythm Coordination and Conversational Structure

A rhythm-based approach is also claimed to be able to account for which intervals between turns are 'salient and potentially informative' (Couper-Kuhlen 1993: 131). That is, those inter-speaker intervals which fall within a rhythmic interval would be less noticeable than those which do not, which in turn could be classed as 'significant pauses'.

Erickson & Shultz (1982) claimed that two parties may integrate the rhythm not only of their verbal interaction, but also of their non-verbal interaction. McClave (1994) has more recently found some evidence to support the gestural rhythm hypothesis. Couper-Kuhlen and Auer found several areas in Erickson & Shultz' work in need of clarification. Two points - what counts as a rhythmic beat? and how may one judge the regularity of beats? - have already been outlined in terms of metrical structure. The third point is of concern here: what (if anything) does rhythm contextualize in verbal interaction?

Conversational interaction is often concerned with two criteria: who the next speaker is, and what topic the next speaker is expected to talk about. Couper-Kuhlen refers to each of these as having two 'values': *determined and non-determined.* We therefore have a 2x4 grid, where each combination may be termed a transition type.

Who next?	**What next?**
1. determined	determined
2. non-determined	determined
3. determined	non-determined
4. non-determined	non-determined

For example, when there are only two interlocutors, and one asks the other for his name, we have a type 1 transition - both the next speaker and topic are determined. A type 4 transition may arise in a group discussion, where a chair may ask 'So, does anyone want to say anything before we begin?', where the next speaker and the new topic are not determined. Note also that transition types 1 and 2 are referred to as sequence internal. That is, the next topic is determined. Transition types 3 and 4, where the next topic is non-determined, are referred to as sequence external.

The significance of this framework for the rhythmic view of turn-taking becomes apparent when one considers that, for example, in a type 4 transition there is little pressure for anyone to say anything in particular. There is, one might suppose, therefore less likelihood that a next speaker will have been noting the rhythmic pattern of a current speaker than in a type 1 transition. Note that although this supposition was applied by Couper-Kuhlen to rhythmic turn-taking mechanisms, it is equally applicable to a linear model, since in a type 4 transition a next speaker may not have been as aware of the normal intonational, pragmatic or syntactic cues indicating an impending TRP. Presumably, if the next topic for discussion is not determined, one might also expect that there would be some degree of hesitation (an 'awkward silence').

Couper-Kuhlen found rhythmic integration at all four types of transition, although its function was different in each case. Most surprising was the apparent existence of rhythmic integration at sequence external junctures - cases where the next topic is non-determined.

Couper-Kuhlen's point is that the variation between rhythmic and arhythmic sequences in conversations can be explained when the discourse structure of the

utterances is taken into account. Following from this, an arhythmic transition need not cue 'awkwardness'. The general possibilities of rhythmic coordination outlined earlier may be contextualized in different ways. For example, with immediate isochronous onset the next speaker's first prosodic prominence coincides with the next beat following the current speaker's TRP. Because immediate isochronous onset is taken as the unmarked case, interlocutors ought to perceive transitions of this type to be the most 'natural'. Whether they do or not is uncertain. Alternatively, a delayed isochronous onset presents no breakdown of rhythm as such. Instead, a beat is 'missed' between the TRP and the first stressed beat of the next speaker's turn. Couper-Kuhlen & Auer (1988) note that more than a one beat delay may occur. No exact upper limit is postulated, although one might imagine that more than a two or three beat delay would be stretching the concept of rhythmic coordination too far.

Cases of immediate isochronous onset followed by rhythmic disintegration, or anisochronous onset could present a potentially serious problem for the conversation, since an established rhythm breaks down. Where no rhythm is established at all, the consequences are quite different, and one might suppose less serious, than the case of rhythmic disintegration following an isochronous onset.

The advancement upon the earlier Erickson & Shultz (1982) model is that the contextualizing function of rhythm in conversation and turn-taking is more than a varying degree of 'comfortableness'. Face threats may be contextualized as an onset which does not fall 'on the beat', yet which begins a turn with its own rhythmic structure. Thinking time - a 'time out for calculation' - may be contextualized as a delayed onset. Here, the next speaker is faced with the conflict between the desire to maintain the rhythmic structure of the current speaker's turn, and the necessity to stop and plan the next turn. Couper-Kuhlen & Auer also rule out such a direct link between rhythm and context as being too deterministic. They claim that, for example, a potential face threat can be controlled by rhythm. They propose that the link is variable, or negotiable, thereby putting greater communicative function on rhythm.

But while these hypothesised communicative functions of rhythm cover neatly different contextual situations in conversation, one wonders whether they offer any great advantages over a linear account. Conversational face threats, 'smoothness'

of conversation, or other unquantified phenomena concerning an attitude to a conversation can surely be described as well in terms of the relative duration of an inter-speaker interval in a given context (i.e. whether there is a question/ answer sequence, whether, as in the map task, one interlocutor typically asked for confirmation from the other).

2.5.8.3 Speech Rhythm Coordination and Timing

One advantage of the rhythm-based view of turn-taking is that it acts as a measuring device by which the onset of a turn in a conversation could be judged as well- or ill-timed. This rhythmically timed interval would not be of any certain duration, but would vary according to context. Couper-Kuhlen (1993) points out that a model of turn-taking based on a linear notion of timing in interaction is unable to account for a number of aspects of turn-taking. For example, according to Sacks et al. turn-taking involves the mutually constraining principles of *earliest possible start* and *intelligibility*. These two principles are claimed to create the tight temporal coordination between conversational turns mentioned earlier. But the logical outcome of this would be that latching[10] would be the unmarked case in turn-taking, which Couper-Kuhlen points out is not the case. She suggests that a rhythm-based view of timing at speaker transitions is the solution to this and other problems with the current model of turn-taking. As Martin (1972) noted, rhythmically patterned sounds can be tracked in such a way that the perception of initial elements of the series allows later ones to be anticipated. The perception of sounds which were merely concatenated would be cognitively more demanding, since continuous attention would be required by the system (cf. the earlier comment about rhythmic 'gestalts').

It seems uncertain why this should be an advantage over the alternative view that turn-taking is timed on a linear basis, with the only constraints being social and cognitive ones. Well- or ill-timed next-speaker onsets could be measured in terms of sociolinguistic norms for the level of overlap, or the maximum amount of elapsed

---

[10] The case where the second speaker begins immediately as the first finishes.

time allowable before next-speaker turn onset. In combination would be the amount of time needed by the next speaker to plan the next utterance.

Likewise, the claim that rhythm eases the perception of speech does not seem to hold for turn-taking. While it may be true within an utterance, there does not seem to be any similar theoretical advantage across turns. The linear view may seem more complex, but this is no reason to suppose that a rhythm-based approach is any more likely to hold, even if it is more advantageous to believe that it may because of its relative simplicity.

2.5.9 Summary

Couper-Kuhlen has claimed that isochronous sequences can be perceived in speech, both within and across conversational turns. She claims that the perception of isochrony is part of a more general psychological 'chunking' process. But also, the presence or absence of perceptual isochrony may act to signal the social relationships between participants in a conversation. The hypothesis is that isochronous sequences indicate that participants share the same interpretive framework. However, there is no evidence either for isochrony as a psychological chunking process, or as a contextualization cue.

Nevertheless, lack of immediate evidence does not refute the hypothesis. The notion of perceived isochrony as a coordinator of turn-taking offers some advantages for an understanding of the organization of conversation. Although worthy of further analysis, empirical evidence for the rhythmic coordination hypothesis remains sparse. Further empirical analysis is the subject of chapter 5 of this thesis.

In the next section I shall cover an alternative theory to the ones already mentioned, which was developed by Clark (1996). The rhythmic coordination hypothesis gives no real account of contextual information, such as the social relationship between two participants in a conversation, or the possible varying cognitive states of the participants. It presumes that perceptual isochrony acts as a coordinator of turn-taking, and that when isochrony is not present at a turn transition, that transition is in some way marked. The model developed by Clark gives an account of coordination based on the context of social and cognitive elements of the

conversation. The timing of coordination can be expressed in terms of these contexts, rather than in terms of isochronous or non-isochronous sequences.

2.6 The Clark Model

In this account conversation is viewed not as a process of turn-taking consisting of ordered allocation rules, but rather as part of a general process of the coordination of joint actions between two or more people. It proposes that conversation should not be treated purely as a cognitive or probabilistic entity in which one person responds to a stimulus, though speaker and listener otherwise act independently. Nor does it propose that conversation should be treated as a purely social-interactive process. Instead it assumes that conversation is a social process of carrying out a joint action using various signals. These signals reveal information to both parties about their respective mental states and level of understanding of the signals, as well as the individuals' social relationships.

2.6.1 Coordination Problems

In Clark's view, conversation is essentially a form of coordination problem (Lewis, 1969). In coordination problems, two or more people have common goals and must reach some solution based purely on what one expects the other to do. In coordination problems, one participant (A) must attempt to make certain that the other (B) has perceived and understood the signal. B must make certain that A is aware that the signal has been understood. In dialogue, the coordination process is continuous, requiring moment-by-moment decisions. Clark (1996) refers to these coordinated acts as *joint actions.*

People use *coordination devices* to allow one person to make assumptions about another person's likely actions based on the known shared, communal or personal, experiences of the pair. For example, a simple coordination problem might be set up by telling two people independently to choose between 'heads' and 'tails', and to write their choices down, with the understanding that if they both were to make the same choice, they would win a prize. Interestingly, Schelling (1960) found

that 86% of people presented with this problem chose 'heads'. The coordination device in this apparently intractable problem is the assumption that people very often make that other people are likely to choose heads, if they share a similar cultural or personal common ground with them. Even in simple problems, people have certain expectations and presumptions which they use to the advantage of all concerned. Coordination devices, then, give participants in a coordination problem a basis for believing that they will converge on the same joint action. Clark (1996) describes this in terms of a *Principle of Joint Salience*:

The ideal solution to a coordination problem among two or more agents is the solution that is most salient, prominent, or conspicuous with respect to their current common ground. (Clark, 1996. p. 67)

Clark also suggests that these joint actions can be coordinated because they are divisible into units called *phases*, which themselves can be divided into sub-phases. Joint actions therefore have a hierarchical structure. Phases form the basis of all joint actions, and it is the phases, or sub-phases, which get coordinated. All phases consist of three parts - an entry point, a body, and an exit point.

When participants coordinate successfully with the entry and exit times of each phase, they are said to be in synchrony. For this to happen, there must be some form of projection of the entry and exit points. Clark suggests that three strategies are used for this projection.

i) *Cadence strategy*. This is only used in periodic activities. Periodic activities are those joint actions which are synchronised by rhythm, such as marching, dancing, or music. Bars are the phases in music, and the entry times for these phases are marked by heavy beats. The duration of a phase is predictable from the rhythmic beat. The rhythmic coordination effectively makes the claim that coordination is achieved through the use of a cadence strategy.

With a cadence strategy, participants coordinate by agreeing on:

a) entry time, t

b) duration, d

c) for all participants i, the participatory action p_i that i is to perform in d

ii) *Entry strategy.* This strategy applies for all continuous actions and is therefore more general than the cadence strategy. Because an exit from one phase in continuous actions means automatic entry into another, coordination is only required on:

a) entry time, t
b) for all participants i, the participatory action p_i that i is to perform in the phase

The problem that lies in this strategy is to project the entry times successfully, since there is no predictable beat that can be used as with the cadence strategy. In fact, in conversation participants appear to use a variety of strategies to project entry times for each new phase. These strategies include the use of intonation, gesture, eye contact, syntax, and pragmatics.

iii) *Boundary strategy.* On occasion, there is neither a cadence nor an immediate and automatic transition from the exit point of one phase to the entry point of the next. Participants need to coordinate on three features:

a) entry time, t
b) exit time, u
c) for all participants i, the participatory action p_i that i is to perform in the phase

For example, if two people were to shake hands, and if the various processes involved in shaking hands were split into phases (extend hands, shake hands, withdraw hands) the final phase of shaking hands is not followed by any other phase. Therefore the exit time of that phase must be projected from the actions of the current phase.

Clark goes on to suggest that there is a basic principle behind these three strategies - *The synchrony principle.*

In joint actions, the participants synchronise their processes mainly by coordinating on the entry times and participating actions for each new phase.
(Clark, 1996. p. 86)

2.6.2 Common Ground

Any joint action is only possible with *common ground* (Stalnaker, 1978). Common ground is the set of mutual knowledge and beliefs shared by two or more people. Part of each participant's self-awareness involves knowing that other people have an analogous self-awareness. According to Lewis (1969):

proposition *p* is common ground for members of a community *C* if and only if:

1. every member of *C* has information that basis *b* holds;
2. *b* indicates to every member of *C* that every member of *C* has information that *b* holds;
3. *b* indicates to members of *C* that *p*.

Common ground can be categorised broadly into two types - *communal common ground* and *personal common ground*. The former is based on assumptions people make about the sorts of cultural communities that other people belong to. If there are shared bases between two or more people at a cultural level, they will have some cultural common ground. For example, if two strangers meet on holiday and discover that they are from the same country, each will immediately make assumptions about the sorts of things the other is likely to know and to have experienced in growing up in that country, based on this cultural common ground.

Personal common ground is based on assumptions about what one person knows about another person. If there are shared bases between two or more people at an individual level then they will have personal common ground. Personal common ground might range from the assumption by one person listening to a concert that

someone in the next seat can also hear the music, to assumptions made between friends about past shared experiences.

Note that the assumptions made in both types of common ground are a matter of degree. Many assumptions can be made with a great degree of certainty (for example, assuming that someone from Britain will know that Tony Blair is the Prime Minister, or assuming that someone standing next to me in the street also heard a loud explosion) whereas others are made with less certainty (for example the assumption that a fellow photographer will have used a dark room before).

Without common ground any form of joint action would not be possible because the people attempting to form the joint action without any common ground would, quite simply, be unable to understand one another. One participant's signals would be meaningless to the other, and vice versa. However, for there to be common ground is not enough. It has to be staked out - a process which is called *grounding* (Clark, 1996).

2.6.3 Grounding

According to Clark, to *ground* a joint action is to establish it as a part of common ground. In conversation, participants work together towards the mutual belief that the signals passed between them have been understood well enough for current purposes. The signals may, of course, have been misconstrued - but it is enough if at any given time both participants each have the belief that they understand the signal and that the other participant realises that the signal has been understood.

It follows from this that each participant needs some form of evidence that their respective actions have been completed, and understood. They require *closure* on their actions. This is required for all forms of action, not just for joint actions. When participants are involved in a joint action, the *Principle of Joint Closure* applies:

The participants in a joint action try to establish the mutual belief that they have succeeded well enough for current purposes. (Clark, 1996. p. 226)

If participants are to believe that they have succeeded with their actions, they have to make the assumption that the evidence available is sufficient for this belief. There are three factors which make this possible:

1. Validity of evidence. The feedback from an action must be valid. If it is not, and cannot be relied upon, then there is no way that the agent performing an action can know whether that action was successfully performed

2. Economy of effort. The easier the feedback from an action is to acquire, the better

3. Timeliness. Feedback from an action must take place within a reasonable time frame. If it does not, then the delays for that phase may transfer to the next phase, causing delays there.

These three factors generally rely on two principles:

1. The Principle of least effort: All things being equal, agents try to minimize their effort in doing what they intend to do.

2. The Principle of opportunistic closure: Agents consider an action complete just as soon as they have evidence sufficient for current purposes that it is complete.

It will be noted that integral to the concept of grounding is time. An action is considered complete as soon as there is enough evidence that it is complete, and response or feedback to that action must take place within a reasonable time frame. I shall therefore turn towards an account of how conversation may be described in terms of joint actions and grounding, and the importance of timing in the coordination of conversation.

2.6.4 Conversation and Joint Actions

Clark describes conversation as composed of *joint actions* (Clark, 1996. Chapter 3).

It therefore consists of phases. In conversation, these phases form a hierarchical structure ranging from utterances to phonetic segments. For conversation to proceed smoothly, the speaker must suppose that the listener has finished processing each phase approximately when the speaker finished producing it.

Although phases in conversation occur at different linguistic levels, the phase which Clark claims to be the most salient is one approximating to the Turn Constructional Unit (TCU). There is no precise definition of this unit, although it would seem to be approximately equivalent to an intonation unit.

Participants in a conversation attempt to ground their actions - to reach a joint closure. A contribution occurs when the joint closure is achieved successfully. Contributions consist of two main phases - presentation and acceptance. The presentation phase consists of the presentation of the utterance itself by one participant to the other. The acceptance phase consists of the other participant communicating what has or has not been understood. What is noteworthy about this approach is that the contributor requires positive feedback from the addressee that the information has been accepted and processed. These positive feedback signals may be divided into four categories:

1. Assertions of understanding. These are assertions that a signal has been understood by the addressee through the use of nods, smiles, or expressions such as 'mm-hmm' (in other words through the use of backchannelling).

2. Presuppositions of understanding. The addressee presupposes that he or she has understood the addressor well enough to continue to develop the conversation as soon as he or she takes up the conversational floor in response to the signal.

3. Displays of understanding. The form that the addressee's contribution takes displays what has been understood by the meaning of the contributor's signal.

4. Exemplifications of understanding. These are similar to displays of understanding, except that the addressee signals correct (or incorrect) construal of the contributor's

signal through iconic gesture, paraphrase, or repetition.

2.6.5 Levels, Tracks and Layers

An important area within the Clark model is the treatment of language as being made up of different lines of actions. These fall within three dimensions: levels, tracks, and layers.

2.6.5.1 Levels

Clark proposes that there are four levels of joint action in every utterance.

Level	Actions
4	C gets N to participate in joint actions
3	C gets N to understand the signal
2	C gets N to identify actual expressions
1	C gets N to attend to the signal

The actions of C and N are linked at all four levels.

2.6.5.2 Tracks

On the one hand, people speak to one another for some reason - for example to get someone else to do something, or to convey facts or information. These actions are in what Clark calls *track 1*. At the same time that the participants are conveying and accepting information in track 1, they are also attempting to form a communicative act. These meta-communicative acts are in *track 2*. In effect, track 2 consists of signals which are used to establish that the signals in track 1 have been properly communicated and understood. They can be real or inferred signals. For example, take the following simple exchange:

	Track 1	Track 2
A:	take your line from the start and go vertically down, past burnt forest	[do you understand this?]
B:	[I'm following your directions]	right [yes, I understand]

But B's signal in track 2 could equally have been composed of silence, or a nod of the head. The important point to note here is that B's 'right' in track 2 effectively acts as a repetition of A's utterance, and confirms that it has been understood. Track 1 is used to signal tacitly that the directions given by A are being followed.

2.6.5.3 Layers

Utterances may consist of two (or more) layers. A simple utterance, which states some fact or simple piece of information, has only one layer of actions. However, many utterances may rely on pretence, or non-literal meaning. For example, novels, jokes, anecdotes, sarcasm, and rhetorical questions rely on a joint pretence by participants that certain events have occurred. This *pretence* occurs in *layer 1*. *Layer 2* consists of the events themselves.

As Clark points out, levels, tracks, and layers are ways of representing language-users' tacit understandings of conversations:

> People tacitly know what it is for listeners to attend to a speaker's utterance without identifying it, or to identify it without understanding it, or to understand it without taking up the speaker's proposal. People tacitly know what speakers are doing when they say "um" or "I mean." People tacitly know what speakers and addressees are doing when speakers tease, become sarcastic, or make ostensible invitations. (Clark, 1996. p.391)

2.6.6 Emergence of Orderliness

Clark's model assumes that the organisation of conversation emerges from the way that people propose and engage in joint activities, and attempt to achieve joint

closure. Clark argues that out of this arise a set of procedures which determine who speaks and acts when. There are three procedures: *minimal joint projects, one primary presentation at a time*, and *presentation and acceptance phases.*

2.6.6.1 Minimal Joint Projects

If C produces an utterance, he or she will expect that N will provide evidence that the utterance has been understood. C will also presume that N's utterance will be a response to it. By making a response, N signals that the utterance has been understood, and that he or she has taken up the joint project. As Clark points out, this accounts for rule 1a in the Sacks et al model, because it explains who is selected next, and why N is expected to begin a new turn upon selection - although N has the right to decline the opportunity.

2.6.6.2 One Primary Presentation at a Time

Participants in a conversation generally cannot concentrate both on presenting and understanding an utterance simultaneously. There is therefore a general expectation that primary presentations will occur one at a time. Note that this applies only to *primary* presentations, and not to secondary presentations (for example, backchannel signals). This procedure allows for the view that the conversational floor is a scarce resource, because each potential contributor can only make a primary presentation while the other participants are silent. Turn-taking is precisely coordinated in order to distribute this scarce resource. Strategies such as overlapping, deliberate interruption, and recycled turn beginnings may be used to gain an opportunity to make a primary presentation.

2.6.6.3 Presentation and Acceptance Phases

Presentation and acceptance phases consist of a presentation of an utterance, where the speaker must attempt to get the listener to understand, and to be able to give evidence of acceptance of the utterance. Clark argues that these phases of

conversation account for the Sacks et al rule 1c - that C may continue speaking until N is selected or self-selects. However, this procedure can account for more than rule 1c, because it can account for observations such as N's backchannelling, N's acceptance of one instalment of a larger utterance, or expansions of an utterance when C realises that N has delayed acceptance.

2.7 Timing

The limitation inherent in all the models discussed so far is that they do not account adequately for certain elements (or for that matter any elements) of timing. It is presumed in both the signalling and sequential-production approaches that N may respond to C's offer to pass over the conversational floor after some unspecified amount of time.

The general assumption must be that N does not passively listen to what is said, and take the floor when C has finished, because this would often not allow N enough time to take the floor. It has been observed that if N were to wait for C to finish the utterance before speaking, response latencies would exceed the observed inter-speaker intervals. In other words, N must be able to *project* the imminent closure of C's turn. The turn-taking rules of Sacks et al. may describe how N is potentially selected by C or how N self-selects, but they do not account for how N knows when C has finished or is about to finish the turn (whether a TRP is likely to occur). The same can be said for the Clark model, where the entry time must be known for successful coordination. While the rhythmic coordination and slot models do incorporate a timing element of sorts, neither accounts fully for projection of an entry point by N. That is, rhythmic-based models allow for prediction of where an entry point could occur, but they do not also account for why N entered when he or she did. The signalling approach uses a system of cues, and therefore covers possible mechanisms by which N can predict when C's turn is likely to close. It is this system of cues which forms the basis of the projection component.

In the next two sections the projection component is discussed. Also given is an account of the importance of timing in the coordination of turn-taking, and the role timing serves in signalling social and cognitive information based on the

common ground between participants in a conversation.

2.7.1 Projection Component

2.7.1.1 TRP Structure

The mechanism for projecting TRPs must depend on how we recognise them. Ford & Thompson (1995) suggest that TRPs consist of one or more of three types of *completion point*. These are points where some form of syntactic, pragmatic, or intonational boundary occurs. The definition of these three boundaries is given below, followed by a discussion of TRP structure, and Ford & Thompson's claim that completion points are used to project TRP location. The important point for current purposes is not so much how N can predict exactly when C will reach a completion point. Rather, it is that both C and N are aware of a completion point once it has been reached, and that both are aware that it constitutes a potential TRP.

Ford & Thompson used as a data base a series of excerpts (totalling about 20 minutes of talk) from two face-to-face, multi-party conversations in American English. The participants knew each other well. There were two research questions:

> a) to what extent is syntactic completion a predictor of turn completion as validated by actual speaker change? If intonation and pragmatics are considered, is the prediction stronger?
>
> b) Where the convergence of syntactic, intonational and pragmatic completion are not associated with speaker change, are crucial interactional factors at work? Can the residue be understood as evidence of the strategic interactional use of a norm?
>
> (Ford & Thompson, 1995. p. 7)

2.7.1.2 Syntactic Completion Points

Ford & Thompson (1995) set out to test the extent to which syntactic completion is a predictor of turn completion, and whether intonation and pragmatics are equally important factors. They judged an utterance as syntactically complete 'if, in its sequential context, it could be interpreted as a single clause, i.e., with an overt or

directly recoverable predicate, without considering intonation.' (p. 8)

Since syntactically complete utterances can always be extended, points of syntactic completion may be incremental. Thus a point of syntactic completion need not have a complete unit between it and the previous completion point. As an example, take the following (where slashes represent syntactic completion points):

20)

V: And his knee was being worn/ - okay/ wait./
It was bent/ that way/

(Ford & Thompson, 1995. p. 9)

The syntactic phrase 'that way' does not constitute a whole unit in itself. But the completion point at the end of it marks the second potential syntactic completion point of the whole phrase 'It was bent that way'. The first potential syntactic completion point is after 'It was bent'.

Ford & Thompson do not claim that a speaker's talk at one point necessarily constitutes an independent grammatical unit, but that a point in the *stream of talk* does. This definition is quite different from the traditional notion of the syntactic clause, and takes into account preceding context. Take the following:

21)

D: I mean it's it's not like wine/ it doesn't taste like wine/ but it's
W: Fermented./
D: White/ and milky/ but it's fermented

(Ford & Thompson, 1995. p. 9)

Here, a syntactic unit for current purposes may run over from one speaker's turn to another's. This perspective on syntactic completion therefore highlights syntactic structure within a dialogue framework. Under this definition cases of N finishing C's utterance seem less like uncooperative interruption and more like a cooperative, interactive process.

2.7.1.3 Intonational Completion Points

Ford & Thompson use the *intonation unit* as the basis of their analysis of the intonational completion point. For the purposes of their research, Ford & Thompson assumed that the intonation unit was an accepted and well-established unit, which could be identified perceptually on the basis of its intonation contour. The intonation unit is essentially the equivalent of the *intonation group* of Cruttenden (1986), and like the intonation group does not have boundaries which can be determined on purely acoustic grounds. Perception of intonation is based on the degree and direction of pitch movement on a stressed syllable, a change in pitch compared with surrounding speech, acceleration of tempo, final lengthening, and pause location (for example, see Cruttenden, 1986). To determine intonation units perceptually can be difficult enough in read speech, but the problems are exacerbated in the sort of conversational speech characterised by pauses, false starts, and hesitations.

Ford & Thompson deemed intonational completion points to be those points where either a sharp rise or fall could be heard at the end of an intonation unit. Syntactic boundaries were disregarded. They point out that while one might suppose that syntax cannot be disregarded in identifying intonational boundaries, studies (e.g. Schuetze-Coburn et al., 1992) have indicated that syntax may not play a vital role in the perception of intonation unit boundaries, and that in fact reliable judgements can be made even where a judge does not know the language being analysed (or where the signal is low-pass filtered).

Examples 22) and 23) demonstrate Ford & Thompson's determination of intonational unit boundaries, where a slash represents a syntactic completion point, a question-mark represents a marked rise, a full stop represents a marked fall, and square brackets indicate overlapped speech:

22)

V: Okay/ this is what t-the problem is/ .
My dad's knee- leg was very bow-legged/ .
It was like thir[teen degrees/]

C: [all his life/ .]

(Ford & Thompson, 1995. p. 11)

23)

J: Well then he got picked/ up/ by Large Marge/ .

W: y'mean just generic fri:ed meat/ ?

(Ford & Thompson, 1995. p. 11)

2.7.1.4 Pragmatic Completion Points

Ford & Thompson define pragmatic units in terms of *conversational action* (e.g. Schiffrin, 1987; Fox, 1987) and intonation. If an utterance can be viewed as a complete conversational action, and if it is bounded by intonation completion points, then it is a pragmatic unit. Pragmatic completion is not the same as intonational completion, because some points of intonational completion do not also involve pragmatic completion. In 24), pragmatic completion is marked with a greater-than sign:

24)

K: It was like the other day/ uh .

(0.2)

Vera (.) was talking/ on the phone/ to her mom/ ?>

C: Mmhm/ .>

K: And uh she got off the pho:ne/ and she was incredibly upset/ ?>

C: Mmhm/ .>

Ford & Thompson (1995. p. 13)

Two forms of pragmatic completion can be distinguished. Local completions occur where the speaker is still continuing with the general topic of conversation, but where opportunities also occur for the listener to signal attention. The example above demonstrates local completion points, because at the end of both of K's utterances the listener would reasonably expect a further development of the conversational topic. Where this does not happen, the pragmatic completion point would be said to be global.

2.7.1.5 Complex Transition Relevance Places (CTRPs)

The major finding of Ford & Thompson's analysis of conversational data was that there was a high degree of coincidence among the three types of completion - pragmatic, syntactic, and intonational.

25)

total intonational completion	433
total pragmatic completion	422
total grammatical completion	798
intonational *and* grammatical	428
grammatical *and* semantic	417
total points of convergence of intonational, grammatical, and semantic completion points	417

(Ford & Thompson, 1995)

25) shows the distributions of the intonational, pragmatic, and syntactic completion points that Ford & Thompson had identified according to the criteria listed in sections 2.7.1.2 - 2.7.1.4. The number of points of convergence of intonational, pragmatic, and syntactic completion in the data (417) is almost the same as the total number of intonational completion points (433) and pragmatic completion points (422). That is, intonational and pragmatic completion points are usually also syntactic completion points, but that the reverse is not true. There are many more syntactic completion points than of either of the other two types.

So, 'intonation and pragmatic completion points select from among the syntactic completion points to form what we will call 'Complex Transition Relevance Places' (CTRPs).' (p. 15) While use of the term CTRP highlights the composite nature of many TRPs, it presumably covers only a sub-group of TRPs which do consist of more than one completion point. Many TRPs may consist of only one completion point, and it therefore seems easier to refer to all transition relevance

places as TRPs rather than CTRPs outwith the current discussion of completion points.

The notion of intonational completion also becomes significant. Ford & Thompson have shown that all pragmatic completion points coincide with intonational completion points by definition (although not necessarily vice versa), and that almost all intonational completion points coincide with syntactic completion points. However, the role of pragmatics becomes decidedly uncertain because all cases of pragmatic completion would appear to coincide with at least one other completion point. In particular, the results seem to suggest that pragmatic completion points always coincide with intonational completion points, whereas intonational completion points may occur in isolation. Also, Ford & Thompson's judgements of pragmatic completion remain intuitive and provisional.

This does not mean that pragmatic completion points do not occur. As outlined in section 2.2, conversational structure may be categorised hierarchically according to *functional* units, and it seems clear that any utterance may be sub-divided to some extent according to the discourse tasks that its parts may carry out. Hence the need for an account of the pragmatic units of an utterance. The problem lies in determining the role that pragmatics may play. This seems to be more of a supportive role, and may act to confirm what N already was able to decide using syntactic and intonational information.

2.7.1.6 TRPs and Speaker Change

TRPs were found by Ford & Thompson to match well with speaker changes. If this were taken as the unmarked case in turn-taking, one would expect that when speaker changes do not coincide with TRPs there would be some interactional force to the speaker change. For example, overlap may be used intentionally to indicate familiarity, either with the speaker or with the information presented, rather than arising from a miscalculation of a projected TRP (see also Jefferson, 1973). Intentional overlap can often be an attempt to complete the other speaker's utterance,

or to indicate attention or agreement.[11] Coulthard (1977) also notes how in some cases N may simply interrupt at any point, which in some cultures is often taken as a sign of impoliteness. Alternatively, desire for the floor may be expressed through a series of short utterances, often just the first word or syllable of the potential next turn. For example, one may often hear a next speaker interrupting with interjections such as '*Bu-, bu-, but...*'

Unintentional overlaps may also occur, typically where N has incorrectly detected a TRP. C has not selected N, but has given sufficient signals that the turn is almost complete. Incorrect self-selection, then, can lead to some amount of overlap, and this is typically remedied as one of the speakers quickly yields the conversational floor. If there are two or more self-selecting next speakers, the first starter often is given the right to continue (Coulthard, 1977). In this respect, it is clear that intentional and unintentional overlap are quite different phenomena - the former does interactional work, whereas the latter is simply an error.

Ford & Thompson found that 31% of the total number of TRPs did not coincide with a speaker change. In some cases, N may not wish to take up the floor when the opportunity arises. In this case, turn-extension is commonly used to elicit a listener's response by signalling or re-signalling completion - in other words to 'project a link' for the listener (Sacks et al., 1974). However, in other cases it appears that a speaker uses various strategies to 'mask' the CTRP. Wilson & Zimmerman (1986) and Schegloff (1982) note that speakers may speak faster as they reach a potential TRP, and Coulthard (1977) points out, speakers will often speak more loudly, in a higher pitch, and more quickly (and hence with less attention to grammar and phonology) in an attempt to avoid interruption. This is a signal of intention to extend the current turn despite pragmatic, syntactic, or intonational evidence to the contrary. The increased rate of speech may not simply act as a mutually recognised signal of intention to continue, but may make it physically harder for N to coordinate an entry point successfully. It is as though by using these linguistic techniques they are 'smoothing over' any possible TRPs, and simply blocking an entry from a

---

[11]Backchanneling may fall under this type of overlap.

potential next speaker by force.

The problem of incorrect self-selection, and of C showing intention to continue beyond a TRP, may be reduced further through what Sacks calls an *utterance incompletor* - examples being words such as 'but', 'and', or other clause initiators. However, Ferguson (1975) noted that 28% of interruptions occurred after such apparent markers of continuation, so while they may signal intention to continue, N can ignore them. The resulting speaker switch would be perceived as an interruption, even though the change took place at a CTRP and even though there may have been no overlap of speech signals. Other syntactic devices may be used. For example, C can use subordinate clauses, where N is aware that at least two clauses will follow (for example, a subordinator such as *'if'* will delay the TRP until after the later occurrence of a clause beginning with 'then'). Even more opaquely, a current speaker could use a device such as 'I have three points to make on that...' Another technique for continuing past a CTRP is the use of direct reference to N's attempt to take the floor. After an interruption, C may say something like 'Please, if you could just let me finish'.

2.7.1.7 Completion Points and TRP Projection

Ford & Thompson's account is useful in making relatively explicit the strategies that may be employed in recognising completion points, and hence potential TRPs. But it is no clearer how N is able to predict with any precision *when* a TRP is going to appear.

Unfortunately, Ford & Thompson provide no data on the exact timing of exchanges such as the completion of another's utterance so we cannot know whether the first and second utterances were latched (the start of the second utterance follows on directly from the end of the first utterance), or whether there was some interval between the two. Anecdotal evidence might suggest that often the completion of others' utterances only takes place where C hesitates *before* reaching a completion point, and N feels under some social obligation to offer a guessed completion for that utterance. Further analysis of utterances completed by others could be highly revealing, because if latching were found in these instances it would indicate that

next-speakers are able to predict a TRP with great accuracy, and even how far in advance the prediction may be made. If some interval were found to exist, then that would give an indication of the speed with which N can react to an opportunity to take the floor outwith the TRP. These timing considerations will be covered in greater detail later in the thesis.

In fact, research on the projection of turn completion has been relatively scarce compared to the extensive literature on discourse analysis generally. Jefferson (1973) used the notion of a possible completion point (PCP)[12] to explain her observation that in some cases of overlap N in fact times the start of a turn with a point which would normally have been the end of C's turn, but which was not because of some added tag sequence. Goodwin (1981) demonstrates the different factors which are used to project the ends of turns and to extend a turn beyond the first point where the turn could potentially change - the TRP.

Other research includes the work of Schegloff (1980, 1982, 1987, 1988) on the 'turn constructional unit' (originally given in Sacks et al., 1974) suggests that turns can be sub-divided into units, and notes strategies that speakers can use to extend their turn across unit boundaries. Goodwin & Goodwin (1987) provide examples which demonstrate that the projectable aspects of a turn are confirmed before the completion of that turn. Listeners are able to use the intonational, semantic and syntactic properties of words to make judgements about the type of utterance before the turn is completed. For example, intensifiers like 'so' in utterances such as '*It was s::o: good*' allow listeners to project that the utterance is an assessment as soon as the intensifier is finished.

Generally, research on this topic (see Jefferson, 1973; Goodwin, 1981; Oreström, 1983; Wilson & Zimmerman, 1986; Davidson, 1984; Local & Kelly, 1986; Lerner, 1987; Levelt, 1989) all concurs with Ford & Thompson's work on the issue of projectability of a TRP based on a combination of prosody, syntax, and

---

[12]It should be noted at this stage that the literature can alternate to some extent between the terms 'possible completion point' and 'transition relevance place'. The latter seems to have found general favour and will be used in the rest of this thesis.

semantics, although once again it should be stressed that there appears to be a lack of a clear distinction between cues used to *project* likely TRP location and cues used to *detect* a TRP once it has occurred - two related but distinct issues. In other words, there is agreement on the sorts of cue which signal a TRP, but it is still unknown how these cues might be signalled and understood such that a TRP can be *projected* successfully. It is also unknown how far in advance of the TRP successful projection may typically occur. It is not enough that N be able to predict and detect a TRP. Coordination of turn-taking requires fine temporal coordination.

2.7.2 Timing

Participants synchronise turn-taking according to entry times, and these entry times are synchronised according to the ability of N to project the end of C's contribution. The timing involved in conversation is extremely precise, and there is usually a relatively small window during which the next speaker may take up an opportunity to hold the conversational floor.

Consideration of timing in conversation is required if a model is to account for how participants coordinate. But timing may also be treated as a signal. Variations in entry times can give insights into the mental states of the participants in a conversation. For example, a delayed entry time might show hesitation or uncertainty on the part of the next speaker. Clark (1996) suggests that early entry times may indicate that the next speaker has miscalculated the end of the current speaker's TCU, or has misunderstood the message. Early entry may also be used as a strategy by the next-speaker to indicate to the current speaker that the message has been fully understood and agreed with.

Clark (1996) proposes that entry times provide information about mental states and processing difficulty, and is based on the *Principle of Processing Time:*

People take it as common ground that mental processes take time, and that extra processes may delay entry into the next phase. (p.89)

In Clark's view, conversation is a form of joint action which occurs in real time, and which moreover would be largely useless if unconstrained by time. Because

conversation is a joint action in time, any actions carried out by one participant must occur in the other's time, making time a scarce resource which has to be distributed and managed in some fashion which is acceptable to both. Clark (1996) refers to this constraint as the *Temporal Imperative*:

In a joint action, the participants must provide a public account for the passage of time in their individual parts of that action. (p. 267)

Linked to this, he argues, is a second imperative. *The Formulation Imperative*:

Speakers cannot present an expression before they have formulated it. (p. 267)

This is a significant point, because delays in formulating an expression may result in an untimely delivery. In fact, this imperative lies at the heart of the *Principle of Processing Time*, in that the presumption is that delayed entry points must be, in part at least, the result of some difficulty in processing.

The two imperatives act as competing pressures. On the one hand, N must make best use of the scarce resource of time, but on the other hand he or she must still be able to understand the previous utterance, and formulate a reply. This is true not only for the entry times of each new phase in the conversation, but also for time within a phase. So, a speaker mid-utterance is constrained as much by time as at the beginning of the utterance when taking the conversational floor from the previous speaker. The ideal case as far as the general flow of conversation is concerned is not to waste time, and therefore to continue speaking without stopping (temporal imperative). The worst case for conversation (although arguably the best in terms of eloquence of speech) would be to stop completely mid-utterance, or to stop for long periods of time, in order to plan in depth the rest of the utterance (formulation imperative).[13] However, some compromise has to be reached, and there must be

[13]There would appear to be exceptions to this, such as the use of mid-utterance pauses for dramatic or rhetorical purposes. Presumably, these pauses can be compensated for by other participants in a conversation, who perhaps expect some form of emphasis from the context of utterance or gesture.

some degree of tolerance of pauses or hesitations mid-utterance. Such tolerances apply, by implication, to the transitions between different speakers' turns at holding the conversational floor. If a hesitation by C in mid-utterance is too great to be tolerated by N, then N will start speaking. Therefore, the tolerance threshold will represent an upper limit of generally acceptable intervals between speakers' turns at holding the floor.

Although this upper pause limit is variable, speakers appear to limit acceptable pauses to approximately one second (Jefferson, 1989; Boomer, 1965). If C hesitates longer than this, or N takes longer than this to take the conversational floor when given an opportunity, it is almost always the case that C will retake the floor.

The notion that timing is related to the mental processing of the participants in a conversation is a fundamental principle underlying the research reported on in this thesis. It arises from research in the area of mental chronometrics, which has shown that there is a relationship between processing and time.

2.7.1.8 Mental Chronometry

The information processing approach to understanding mind attempts to isolate basic mental operations, and to understand their relation to subjective experiences and brain processes (Posner, 1978). He goes on to say that:

> The assumption that mental operations can be measured in terms of the time they require is fundamental to modern cognitive psychology. It has provided an objective tool for the systematic observation of mental events, whether or not they are conscious. (p.7)

He defines chronometry as 'the study of the time course of information processing in the human nervous system' (p.7).

Chronometry is realised in different forms of time measurement. A common method is to measure the time between a stimulus and a response to that stimulus - the *reaction time*. Another method is to provide a cue, and measure the time taken

before the reaction to a following event reaches its minimum, and therefore acts as a measure of the time taken to encode a cue optimally - the *encoding time*. Alternatively, it is possible to measure the number of errors made by subjects at given stimulus exposure durations.

Research has shown a relationship between reaction time and the amount of information processing required by a task. For example, it takes 20ms longer to verify that 4 + 3 = 7 than to verify that 4 + 2 = 6 (Parkman & Groen, 1971). To deny that 4 + 5 = 20 takes 50ms longer than to deny that 4 + 5 = 15, apparently because 20 is the correct product of 4 and 5 (Winkelman & Schmidt, 1974). Early work by, for example, Hyman (1953) supported the idea that reaction time would increase linearly with the degree of information processing required. However, later research (e.g. Fitts, 1964) revealed that the relationship was more complex, and depended on the task that subjects were asked to perform, and the type of stimulus presented.

Other experiments relating reaction time with information processing are numerous. In the area of visual processing, Shepard & Metzler (1971) found that if subjects were presented with pairs of shapes, and were asked to decide whether the shapes were the same by rotating them mentally, mean reaction time increased as the orientation of the shapes approached 180° from each other. Cooper & Shepard (1973) found similar results when subjects were presented with rotated letters. Rogers (1974)[14] found that if subjects were asked to say whether pairs of visual tokens were different, their reaction times increased with the number of different features present in the pairs of tokens. Studies into decision tasks involving pairs of letter strings (both real words and nonsense strings) show that mean reaction times to decide that two tokens are the same increase as the number of letters increases (Eichelman, 1970; Beller, 1970).

Clark (1996) lists some heuristics which conversationalists may use for estimating processing difficulty. He suggests that processing should take longer in speaking, all else being equal,

a) the rarer the expression

---

[14]Cited in Posner (1978).

b) the longer the expression

c) the more complex the syntax or morphology

d) the more precise the message

e) the more uncertain a speaker is about what he or she wants to say

Processing should take longer in understanding, all else being equal,

a) the rarer the expression

b) the longer the expression

c) the more complex the syntax or morphology

d) the more precise the message

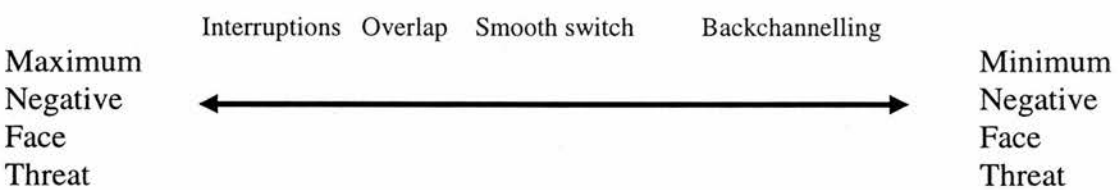
e) the more extensive the implications

f) the less salient the referents

2.7.2.2 Face Threat

Inter-speaker interval durations may also be linked to *negative face* (e.g. Brown & Levinson, 1987). Negative face may be described as a speaker's desire not to be impeded or interrupted during an utterance. An attempt to interrupt an utterance is therefore a potential threat to this negative face.

26)



26) shows how there is a continuum from maximum negative face threat to minimum negative face threat, where interruptions pose the maximum negative face threat, and backchannel signals pose the minimum negative face threat. Backchannels are least threatening because, even though they may sometimes overlap with other utterances by considerable amount, they do not make an attempt to take the conversational floor. In fact, one might argue that they offer positive support, rather than any threat.

In this way, some measure can be given to the relationship between inter-

speaker intervals and social context. Generally, one might say that the greater the overlap, the more negative threat is posed. But this is only descriptive, and not predictive, because face threat is measured in terms of overlap, rather than the other way round. The content of an utterance must be taken into account - whether there is a disruptive interruption, a constructive interruption, an accidental overlap, or a backchannel signal. Content can be described in terms of topic-change, disagreement, agreement, and support (Malam, 1996). Unfortunately, these terms are not clearly defined, but there is at least the potential for them to be clarified and measured, and linked to a notion of face threat, where support provides the least threat, and disagreement provides the greatest threat. A form of predictive link (albeit indirect) can therefore be established between inter-speaker intervals and a notion of face threat. It would be expected that the greater the level of agreement between two participants, the less the negative face threat, and the more likely it would be that there are smooth speaker switches (positive inter-speaker intervals), or inobtrusive backchannelling.

If the context of a conversation, or of an exchange or utterance within that conversation, is known then certain predictions can be made about the relative inter-speaker intervals which should hold in those contexts.

2.8 Contextual Variables in the Coordination of Turn-Taking

By default, it is expected that N should enter immediately, *or nearly so*, after one phase has ended. The extent to which a permissible delay or overlap is tolerated by participants in a conversation is governed by the context of the conversation or exchanges within conversations. This context may be split broadly into two categories - communicative and cognitive.[15]

First, there is what will be termed here the *cognitive context*. This reflects the amount of information processing required by the participants to plan, produce, and

---

[15]The terms 'communicative' and 'cognitive' can be used to cover a large range of concepts. They are used in this thesis in relatively narrow senses to refer generally to mental processing and to relationships between participants respectively.

understand utterances. Work from chronometrics allows the amount of mental information processing required to be defined in terms of time - the more processing, the greater the amount of time required to carry out that processing. The variables which affect the cognitive context are referred to from here on as *cognitive variables.*

Second, there is what will be termed here the *communicative context.* This reflects the common ground that two participants perceive to hold between them. This may include notions such as face threat, and familiarity, as well as the formality/informality of the conversation. Variables which affect social context are referred to here as *communicative variables.*

When timing falls without the bounds acceptable to the participants, the early or late entry time reveals something to C about the intentions of N. In this respect, either intentional or unintentional mis-timing can be regarded as a communicative device.

Work from chronometrics allows predictions about the timing of turn-taking to be made. However, some caution is required to avoid a circular argument - the definition of the amount of processing required in a given context may be based on time, and yet the inter-speaker interval predicted for a certain context depends on the amount of processing involved. It would be meaningless to assume that a query-w move places a relatively high information processing requirement on a listener because query-w moves are generally followed by a relatively long inter-speaker interval, and then to make the 'prediction' that query-w moves are likely to be followed by relatively long inter-speaker intervals. Therefore, only quite general predictions can be made.

An analysis of the transaction, move, and game coding provides some means of accounting for the likely degree of cognitive complexity involved in a sub-section of a conversation. It is to be presumed that different moves, or combinations of moves, generally require different levels of cognitive processing. For example, it would be expected that the level of processing would be less when a speaker N uses an *acknowledge* move in response to an *instruct* move by C, than when N uses a *reply-w* move in response to C's *query-w* move. This is because an *acknowledge* move requires only that N understands C's utterance, and produces some minimal

(backchannelled) response to signal this understanding. A *reply-w* move requires not only that N understands an utterance, but that he or she must also plan and produce a response to a question. This categorisation is only an approximation, because the planning and processing required for, say, a *reply-w* move varies according to context. The prediction from game and transaction coding is that whenever a new sub-task is started, some extra degree of cognitive processing will be required by the participants, because the new task must be planned to some extent in advance. Negation appears to require more processing than affirmation (see the results of Winkelman & Schmidt, 1974), so that one would expect negative responses to query-yn moves to be preceded by a longer inter-speaker interval than positive responses. Following from research on gaze (Anderson et al., 1997), which demonstrated that gaze can be used at moments in a conversation where there is some difficulty in understanding a partner or where there is some more general problem, one would expect that gaze between participants would also be accompanied by greater inter-speaker intervals than when there is no gaze.

A linear, rather than a rhythmic, temporal view of conversation assumes that any intervals or overlaps between speakers' utterances are not governed or constrained in any (significant) way by a supposed rhythmic structure within and between utterances. We have seen that while the use of rhythm to explain social and/or cognitive 'coherence' between participants in a conversation has certain theoretical advantages, there is as yet little evidence that a) rhythm is used as a contextualization cue or b) that perceptual isochrony is used to coordinate turn-taking. If the perception of isochrony exists then it may only be after trained analysts have listened to conversations several times, in an attempt to find it. A real test of isochrony requires untrained subjects, and decisions on an appropriate time-scale for constructing conversations. Normal speakers can judge if utterances or series of utterances appear isochronous, or more 'natural' or 'well-formed'. If the notion of contextualization cues noted earlier were to hold true then presumably perceptually isochronous sequences would appear less 'awkward' than non-rhythmic sequences).

2.9 The Backchannel

At this point it is also worth mentioning one aspect of conversation which is particularly important because of its implications both for a system of turn-allocation rules, and for the projection and timing aspects of turn-taking. This is the use of the *back-channel* (Yngve, 1970). The back channel consists of gestures and speech which do not in themselves constitute a turn. They are used by a listener to signal that he or she is paying attention, and has understood what is being said. Typical back channel signals consist of nods of the head and short unstressed utterances like 'mm-hmm', or 'yeah'.

The problem with backchannelling, as with timing, is that of the 'traditional' models of turn-taking, none can account for it adequately. Such models are generally preoccupied with main channels of conversation. This is particularly evident in the Sacks et al. model, which forces one to assume that backchannelling must be a form of turn-taking. In this section I shall consider backchannelling within different frameworks.

2.9.1 Turns and Backchannels

There is some disagreement about which utterances count as turns and which count as backchannel signals. A broad definition comes from Duncan (1973), who notes that as well as the head nod, 'mm-hmm', and 'yeah' signals, three other types of back-channel signal may be observed:

a) cases where the listener completes C's utterances;
b) requests for clarification; and
c) repetitions of C's utterances or parts of C's utterances

Arguably, some of the types of back-channel signal that Duncan outlined may be said to count as turns in their own right. For example, there seems to be no clear reason why a request for clarification does not in itself constitute a turn, and would be subject to all the conditions of N selection covered by the Sacks et al. model. One factor which determines the status of an utterance may be its length. So, a relatively

short utterance - for example, a one word utterance or a short one or two word repetition of part of another's utterance - might be thought to be operating merely within the back-channel, whereas a longer utterance might constitute a turn. This is vague and subjective, and serves to highlight the problems with accounting for backchannelling within a system which is primarily designed to cope only with main channels.

A more theoretically consistent view of backchannelling is offered by Clark's hypothesis. According to this, talk consists of two parallel tracks of actions (see section 2.6.5.2). Track 2 acts as a medium through which positive feedback signals can be relayed to the other participant, letting them know that their original signal in track 1 has been understood. In effect track 2 acts as a backchannel.

In Clark's view, participants in a conversation make contributions. These contributions normally conclude as soon as a contribution is started by the other participant. However, some forms of contributions are not immediately completed, but continue as a form of general feedback, or backchannelling. If A holds the conversational floor, B may still make contributions to signal to A that the message has been understood, while not actually taking the floor away from A. This process of continued contributions makes use of the backchannel, and the range of signals used in this channel tend to involve short, unstressed, low amplitude utterances such as 'mm-hmm', 'right', or 'OK' in Standard English, nods, or facial gestures.

The main feature of continuing contributions such as these is that they are essentially non-competitive, and therefore one would expect that their location relative to A's utterance or utterances (or, more precisely, to the TCU's within A's utterance) would be of less significance than if B's contribution were to compete directly for the floor. Recall that the principle of processing time allows the prediction that variations in the entry time of B's contribution gives insights into B's mental processes. This essentially arises from the combination of *the principle of opportunistic closure* (agents consider an action complete just as soon as they have evidence sufficient for current purposes that it is complete (Clark, 1996. p. 224)) and the need for timeliness in the closure of any action. In conversation the 'default' case for the entry time of a new phase by B is as soon as possible after B believes - to a

sufficient level of confidence for current purposes - that A's phase has reached its exit point. Note that the latter phrase *sufficient for current purposes* is important, since it makes reference to the common ground between the two participants in the current situation. Common ground will presumably affect the participants' mutual assumptions about what constitutes an acceptable definition of *as soon as possible after B believes A's phase is complete.*

If B's entry time does not therefore conform to the default case, the variation from the default requires interpretation by A. The entry time acts as a signal to A of B's likely mental processes and intentions, given the two participants' common ground. One would expect that a non-competitive positive feedback signal would be recognised as such by A, and that it would be acceptable if that signal overlapped with A's signal. In other circumstances, where B's contribution is a competitive attempt to take the conversational floor, and is recognised as such by A, other construals may be made. For example, that B has incorrectly projected the end of A's contribution, or that B has heard enough and wishes to take the floor early.

2.10 Summary

Under traditional accounts of conversation, two different approaches exist. One approach comes from a 'cognitive' viewpoint, and assumes that conversation is essentially a stimulus-response process, where the current speaker, C, sends out certain signals to the next speaker, N, which indicate when C's contribution is finished. This approach therefore completely ignores any interactive, social element in conversation. Speaker and listener are regarded as isolated entities.

A different approach (for example the Sacks et al. model) is to treat conversation as a social process in which turns are allocated by a set of clearly defined rules. Transitions from one speaker to the next are presumed to be the result of an interactive process, and so this approach views participants in a conversation more as members of a group than as individuals. The advantage of this approach is that it views turn-taking as a projective process - N predicts when C is likely to finish his or her turn. Evidence from analyses of intervals between speakers suggests that very often N will start a turn too soon after the end of C's turn for the timing

mechanism to be based on response to the end of that turn (see for example Couper-Kuhlen, 1993). However, the Sacks et al model makes no explicit mention of the cognitive factors that are involved in conversation, such as the different degrees of cognitive processing that are required in different contexts. It treats conversation as a purely social activity. Moreover, it ignores context in general, and presumes that whatever the situation that a conversation occurs in, the same set of rules will govern the coordination of turn-taking.

Further, a disadvantage with both these approaches is that they assume a rigid form of turn-taking. Conversation may sometimes follow a turnA-turnB pattern, but very often it does not. Speakers may speak at the same time, particularly when one participant in a conversation uses the backchannel. The Sacks et al turn allocation rules also predict that certain phenomena, such as deliberately leaving an utterance incomplete, should be considered as violations of the rules of turn-taking. But these phenomena are usually considered by participants as a normal part of conversation (Clark, 1996).

What is needed, therefore, is a model which can incorporate the cognitive/signalling point of view as well as the interactive approach, and can place them within a framework which does not presume a rigid structure where participants take 'shots' at holding a competitive conversational floor, one after the other.

The model proposed by Clark (1996) comes much closer than other models to meeting these criteria. Clark's central assumptions deal with social, cognitive, and temporal factors:

1. Language is used to carry out social functions
2. Language is a form of joint action
3. The study of language must involve a consideration of social and cognitive elements. That is, conversation is a form of 'social dance' which relies on the social context and social interaction of the participants. But the mental states and thought processes involved in carrying out a conversational task (or any other task) are interwoven with the social interactive processes.
4. Language use is multi-layered

Clark's model is particularly useful because it treats conversation as one form of a more general process of joint actions between two or more people. He argues that the rules suggested by Sacks et al are illusory. Participants appear to be following a set of rules, but in fact they are attempting to take part in locally organised joint projects. Joint actions require a great deal of precise temporal coordination. The null hypothesis is that temporal coordination is not governed by any process than individual preference, and that the distribution of inter-speaker intervals is random. If this is shown not to be the case (and the precision of much temporal coordination, as noted by Beattie (1983), Couper-Kuhlen (1993), and Clark (1996) amongst others, suggests that it is not the case), then the alternative hypothesis is that temporal coordination is context dependent. I have proposed here that there are two likely means by which the coordination could take place.

One is through a rhythmic process, similar to that proposed in the rhythmic coordination model (although that was based within a traditional Sacks et al turn-allocation framework, its basic principles could be extended to the Clark model). The prediction with this model would be that by default the first prominent syllable of N's contribution should fall on or about one of the beats set up by the perceived isochrony in C's contribution. If this were not the case, then according to Clark's theory, the exchange would be socially and/or cognitively marked. N would be able to project the end of C's contribution using intonation, syntax, gaze, and other cues. The rhythmic coordination hypothesis has certain advantages. If correct, it could neatly account not only for observed temporal coordination between participants in a conversation, but it could also act as a framework within which rhythmically or non-rhythmically structured turn-transitions are related to social and cognitive context. However, the empirical evidence in support of this hypothesis is both scarce and uncertain.

A second means of coordinating joint actions may be through simple temporal coordination, where N is expected to make a contribution within a certain time limit after C's contribution - neither too early nor too late. N is effectively

allowed greater freedom in the unmarked case of temporal coordination. If N's contribution does not start within an appropriate temporal window, again Clark's model predicts that it should act as a signal of N's mental state whether in respect of his own contribution or of C's.

Whichever model is used, a central issue (which has yet to be resolved fully) is precisely how N can predict when a TRP is imminent. There appear to be several cues to TRPs, or *turn signals* as Duncan (1974) calls them. Some give more direct cues to closure - particularly syntactic or intonational cues. Body motion and gaze, on the other hand, seem to have less of a direct impact, but do qualify the signals given out in the form of syntax or intonation. As Duncan (1974) observed, the more cues there are, the more likely it is that a change will take place. It seems that a system of turn signals would not be sufficient in its own right as a model of turn-taking, and that the Sacks et al. and the turn signal models should be used in conjunction as separate but related components of the same model.

In conclusion, one possible model might consist of the following components:

1) Coordination component

Joint actions need to be coordinated. That is, the entry point of one contribution should follow as soon as is reasonable after the end of the previous contribution. In conversation, coordination is precise and even slight miscalculations of coordination may not be tolerated. Coordination, according to one model, is to be achieved through the use of rhythmic beats. According to an alternative view, coordination is achieved through a simple temporal process. Whichever approach is adopted, Clark's model predicts that timing information can be used to interpret the social and cognitive states of the participants in the conversation.

2) Communicative component

It is proposed that the theoretical basis of a conversational model must include conversation as one form of a more general class of joint, interactive human activities. Under this view, the study of language and conversation must involve an account not only of the individuals involved, but it must also involve an account of the way these individuals interact, and what it is that they are attempting to *do* in a conversation. The model must include an account of the cultural and individual common ground between the individuals involved in a conversation, since the timing and coordination of speaker switches is governed in part by this.

The view taken here is that communicative variables act globally and locally within a conversation. The global variables, such as participant familiarity or gender, may constrain the general limits (a 'temporal window') within which participants expect inter-speaker intervals to fall. In this respect, global variables only indirectly affect participants' expectations of a tolerable inter-speaker interval for a particular exchange. Local variables, such as the need to signal agreement, acceptance, or disagreement, are also understood by participants to affect inter-speaker intervals of particular exchanges. That is, if a next-speaker wishes to signal disagreement, it is understood by him, and by the other participant, that this can be achieved through overlapping utterances (and possibly starting at a point other than at a TRP).

3) Cognitive component

Likewise, likely cognitive states of participants need to be accounted for. The prediction is that in situations where N is required to carry out more cognitive processing (for example, when answering complex questions, or encountering new information) the inter-speaker interval will be greater (cf. the principles of processing time). The notion of common ground predicts also that C should be able to realise that extra time is required by N, and should allow him or her more time to make a contribution. Cognitive variables operate locally, and affect mutual expectations of inter-speaker interval durations for particular exchanges. They therefore operate within the general constraints of global social variables.

4) Projection component

For N to calculate a suitable and timely entry point also requires projection of a likely closure of C's contribution. Although the exact means by which this is achieved are not certain, it appears that participants are able to use intonational, syntactic, and gestural information to project a transition relevance place.

The next chapter deals with the basic data set used for the analyses reported on in the remainder of this thesis, and the data reduction applied to this data set.

3. A Description of the Map Task Corpus

3.1 Introduction

The dialogues used as the basis for this research were taken from the HCRC Map Task Corpus, which is described in Anderson et al. (1991). The account given of the Map Task Corpus in this chapter is taken from that paper. It is a corpus of spontaneous task-oriented dialogues, based on the Map Task paradigm (Brown et al., 1983), where two participants collaborated to achieve a specific goal. The two participants sat opposite one another, each with a map, which the other participant could not see. The maps included a start-point, a finish-point, and several landmarks (or features). The start-point was labelled on both maps, and all the features were represented with labelled drawings. One participant - the instruction giver - had a route marked on his/her map. The other participant - the information follower - had no such route. The follower also had a slightly different map from the giver, and both participants were told this before the task began.

The participants were told that they had to reproduce the giver's route on the follower's map. The task was for the giver to guide the follower from the start point on the map, to the finish point, following the route marked on the giver's map.

3.2 Materials

Four basic plans were used in the map design, where each plan was devised so that all would give routes of approximately equal complexity. Four pairs of maps (one for the giver, one for the follower) were devised for each of the four basic plans. This gave 16 map pairs in total. Apart from one having a route marked, while the other did not, the maps in each pair differed in half the features critical to the route.

3.2.1 Phonological Characteristics

The maps were designed using four phonological modification categories, or *reduction types*: t-deletion, glottalisation, d-deletion, and nasal assimilation. These reduction types were associated with different feature names.

3.2.1.1 Master Features

Each map contains a potential pair of *master features*, or landmarks, which occurs in one of four contrast/match conditions determining whether one or both of the features is present. Each reduction type is associated with a specific master feature, and each pair of master features appears on an equal number of maps in the corpus. Thus on each map there will be one or both members of a master feature pair taken from the following set:

Table 1

Code	*Reduction-type*	*Master feature names*
1	t-deletion	east lake / west lake
2	glottalisation	white mountain / slate mountain
3	d-deletion	diamond mine / gold mine
4	nasal assimilation	crane bay / green bay

Over the full design each reduction type combines with each contrast/match condition four times.

3.2.1.2 Other Features

In addition to master features, the four categories of reduction type occur as other feature types on the maps, described in more detail below. Each map contains at least one example of each reduction type.

3.2.2 Feature Types

3.2.2.1 Introduction

The maps consisted of a number of landmarks, arranged systematically on a sheet of A3 paper. The giver and follower's maps were constructed to include features which differed along a number of dimensions*: contrast, sharedness, odd-man-out*. These are described more fully below. Additional incidental features were included for lexical variety.

3.2.2.2 Contrast and Match

Over the maps in the design, the pairs of master features appear in a balanced set of contrast conditions. Contrast is a binary variable, and a map is +Contrast when both members of the master feature pair are on the instruction giver's map. There is a contrast in the names of two master features for that map (e.g. *east* lake, and *west* lake). When there is no such contrast, a map is -Contrast. Match is also a binary variable. A map is +Match when the contrast value of the giver's map is the same as that of the follower's map. Either giver, follower, both or neither may have the pair of master features.

3.2.2.3 Sharedness

Along with the contrast/match variable, which defines the character of map pairs, there are classes of features which are defined by the differences between the two members of any given map pair. All the sharedness types occur somewhere in the design with landmark names containing all four reduction types. Examples of the following categories of sharedness were included in each map pair.

Common feature - a feature which has the same drawing, and the same name in the same location on both the instruction giver and follower's maps.

Name change - a feature that is common to both maps but which has a different name on the two maps. For example, the giver may have 'white water', whereas the follower might have 'rapids'. The drawing and location of both would be the same on both maps, however.

Absent/Present - a feature that is present on one speaker's map but not the other's.

Two-to-One (2:1) - a feature of which the giver has two. One of these is relevant to the route and the other irrelevant. The follower has only the irrelevant feature.

3.2.2.4 Odd-Man-Out

Features on a map generally fit a single *scenario*. For example, there might be a "*Wild West*" scenario, with "*Apache camp*", "*canoes*", "*buffalo*", "*gold mine*" and "*cavalry*" as landmarks. One odd-man-out feature would be alien to this scenario, such as a "*nuclear test site*" occurring on the "*Wild West*" map.

3.2.3 Routes

3.2.3.1 Description

Four different routes were constructed for the maps. To help ensure that the routes were different from each other, random number tables were used to generate the co-ordinate points for locating the major features on a basic map pair. A route was then drawn around the features observing the following criteria:

i) the route starts at a shared, or common, feature
ii) the route finishes at a common feature
iii) intermediate landmarks along the route alternate between common features and those that differ in some way
iv) there are at least two features which appear only on the giver's map, and two features which appear only on the follower's.

3.2.3.2 Routes and Master Features

The four routes are associated with particular master features. For example, the master feature "east lake" always occurs in the same location on any map in which it appears, and these maps all share the same route. For this reason a route can be assigned the number given to the phonological reduction type of its master feature. Thus the route associated with the master feature "east lake" is route number 1 as "east lake" is the t-deletion master feature. Routes were assigned as follows:

Route 1 = associated with t-deletion master feature
Route 2 = associated with glottalization master feature
Route 3 = associated with d-deletion master feature
Route 4 = associated with nasal assimilation master feature

3.2.4 Quartets

Each of the 16 maps constructed in the manner described above has a unique route x contrast/match combination. Four different *quartets* of maps were created using a Latin square on this combination. The 16 maps (4 for each quartet) were allocated as in Table 2 overleaf.

Quartet	*Map*			
Qrt1	++1	+-2	-+3	--4
Qrt2	++4	+-1	-+2	--3
Qrt3	++3	+-4	-+1	--2
Qrt4	++2	+-3	-+4	--1

Table 2 - The maps used in each quartet. Each map is represented by a 3-character code. The first two characters represent ±Contrast and ±Match respectively. The third character represents the route number.

+Contrast signifies that the instruction giver's map contains contrasting master features (e.g. both "east lake" and "west lake"). -Contrast signifies that the instruction giver's map contains only one member of the master feature pair (e.g. "east lake"). The presence of +Match means that the instruction follower's map matches giver's in contrast, so if the giver has both lakes, so does the follower. If the giver has only one then the follower has only one. -Match means that the instruction follower's map mismatches the giver's in contrast. If the giver has two lakes, the follower has one. If the giver has one, the follower has two.

3.2.5 Assignment of Feature Names to Feature Types, Maps, and Quartets

Table 3 overleaf shows the assignment of feature names to feature types and maps for quartet one.

Map (master features)	**Type of Sharedness**				
	2:1	*Absent/ Present*	*Name Change*	*Common*	*Odd Man-Out*
++1 east lake west lake	1 fenced meadow	2 picket fence	3 old mill/ mill wheel	4 caravan park	1 nuclear test site
+-2 white mountain slate mountain	4 site of plane crash	3 round rocks	2 hot wells/ hot springs	1 collapsed shelter	4 roman baths
-+3 diamond mine gold mine	3 carved wooden pole	4 saloon bar	1 fast flow-ing river/ fast running creek	2 flat rocks	3 walled city
--4 crane bay green bay	2 wheat fields	1 forest stream	4 cliffs/ sandstone cliffs	3 old lighthouse	2 rocket warehouse

Table 3 - Feature names and feature types for Quartet 1

3.2.6 Examples of Maps

Figures 1 and 2 show the maps given to the giver and follower respectively in one of the Map Task dialogues. In fact, this map was used eight times - twice in q1ec3, twice in q1nc3, twice in q5ec3, and twice in q5nc3. It should be noted here also that dialogues are referred to in the Map Task in terms of their quad number (q), ±eye

contact (e=+eye contact, n=-eyecontact), and conversation number (c). Therefore, q1ec3 refers to the dialogue in conversation 3, quad 1, in the +eye contact condition.

This map is -+3 (-contrast, +match, route no. 3). These show more clearly the differences in features mentioned in table 2. The master features are *diamond mine* and *gold mine.* Note, however, that only the *diamond mine* appears on the two maps. This is because the map is -contrast and +match. -Contrast means that only one master feature appears on the giver's map. +Match means that the follower's map matches the giver's in contrast, and likewise has only one master feature.

The giver's map has two *carved wooden pole*'s, but the follower's map has only one. So, the 2:1 feature is the *carved wooden pole*. The *saloon bar* was present on the follower's map, but not on the giver's, so this was the absent/present feature. *Fast flowing river* on the giver's map became *fast running creek* on the follower's, so this item was the name change feature. The common feature was (amongst others) *flat rocks*. Finally, the map had a 'western' theme, making the *walled city* the odd-man-out.

Figure 1 - The giver's version of map -+3

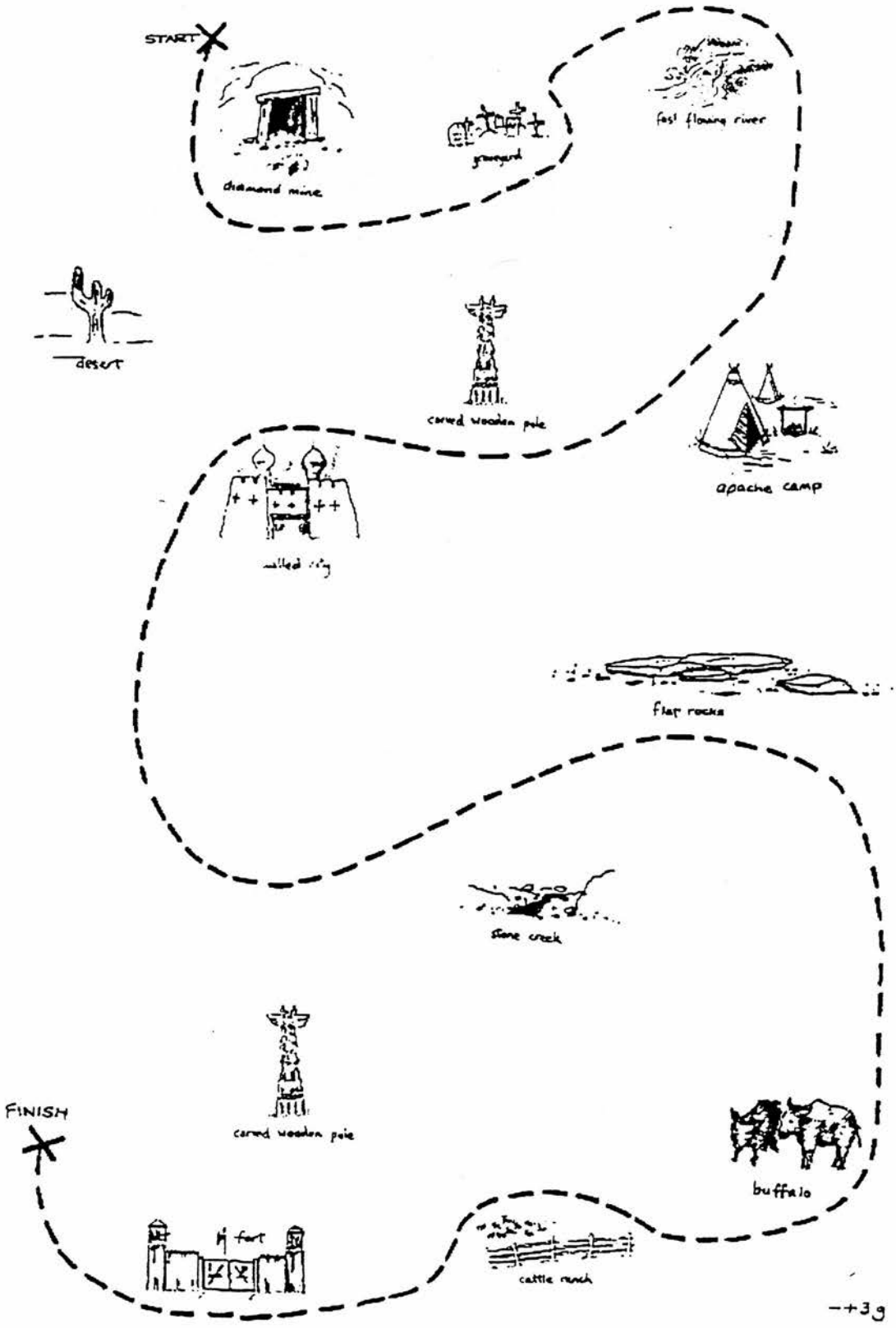
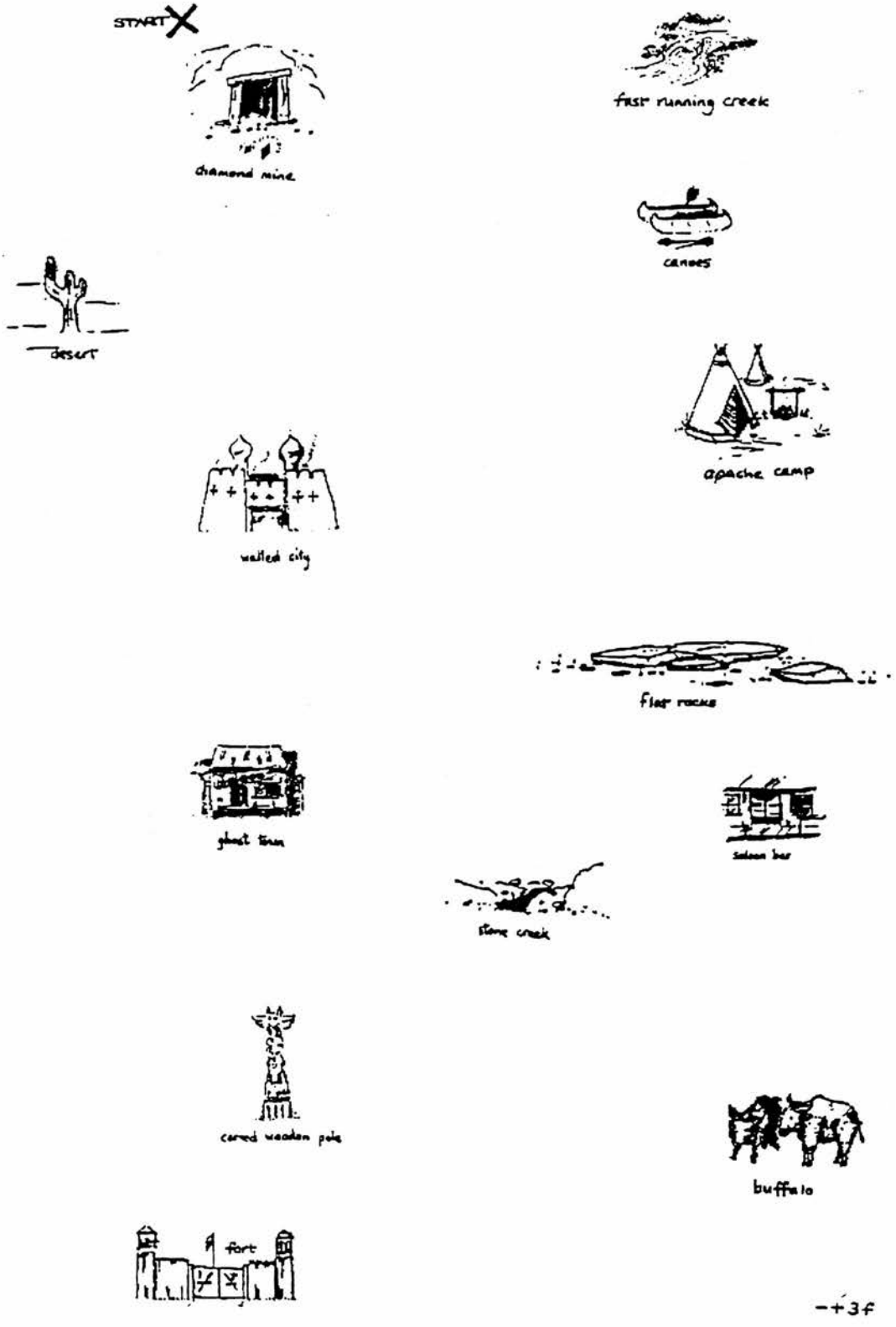


Figure 2 - The follower's version of map -+3



3.2.7 Subjects

32 male and 32 female undergraduates of the University of Glasgow took part in the Map Task. Their ages ranged from 17 to 30, with a mean of 20 years. 61 of the subjects were Scottish. All 64 subjects were native speakers of English.

3.2.8 Familiarity

64 subjects were recruited with a partner that they knew well, giving 32 pairs of subjects. The length of time that familiar partners had known each other ranged from six months to a lifetime, with an average of two years. Of the 32 pairs of subjects, 13 were all-female, 13 were all-male, and 6 were male-female.

Each pair was grouped with another such pair, so that all four speakers formed a quadruple (called a *quad*). Each pair had never before met the other pair. Each quad was assigned a map-pair from each of the four basic map types mentioned above, by Latin Square.

Each subject took part in four conversations. Twice this was as the instruction giver, and twice as follower. In each case as giver or follower each subject was paired once with a familiar subject, and once with an unfamiliar subject. In half of the quads (layer 1) all subjects worked with an unfamiliar partner in their first conversation. In the other half (layer 2) subjects worked initially with a familiar partner. In both layers familiarity alternated, so that if on the first conversation a subject was paired with a familiar partner, on the second conversation that subject would be paired with an unfamiliar partner, and so forth. This meant that each quad consisted of eight conversations, arranged as in Table 4 following.

Layer 1

Conversation No.	*Familiarity*	*Giver*	*Follower*	*Map*
1	-	a1	b1	1
2	-	b2	a2	2
3	+	a2	a1	3
4	+	b1	b2	4
5	-	a2	b2	3
6	-	b1	a1	4
7	+	a1	a2	1
8	+	b2	b1	2

Layer 2

Conversation No.	*Familiarity*	*Giver*	*Follower*	*Map*
1	+	a1	a2	1
2	+	b2	b1	2
3	-	a2	b2	3
4	-	b1	a1	4
5	+	a2	a1	3
6	+	b1	b2	4
7	-	a1	b1	1
8	-	b2	a2	2

Table 4 - Basic design grid, showing the arrangement of familiar pairs of speakers, and map number, for each conversation. a1 and a2 are members of the same familiar pair, and b1 and b2 are members of the other pair.

It can be seen that in each quad each map was used twice. Each speaker also encountered three different maps - one was encountered twice as the giver, and two others as follower. Note that map assignment was identical in the two layers, so that the only difference between layer 1 and layer 2 lay in the different distribution of the familiarity condition over successive conversations.

3.2.9 Eye contact

The 32 subject-pairs were split into two groups - those with eye contact, and those without. In the eye contact condition, the two participants could see each other. In the no eye contact condition, a piece of card was placed between the two participants so that they could not see one another. Apart from this difference, the two eye contact conditions were exact replications.

3.3 Data Files

The data files used in the analyses reported on in this thesis use SGML (Standard General Markup Language) annotation. The reader is referred to Appendix A for an outline of the code used to define the different levels of discourse unit represented in the data.

4. Data Description and Reduction

4.1 Introduction

In this study, my aim was to measure the intervals between speakers' contributions to a conversation *inter-speaker intervals*. A basic problem was therefore to set up a working definition of 'contribution'. Essentially, I used move units and the idea of speaker switches to define a basic unit of analysis - an *utterance*. An utterance by one speaker followed by an utterance by a different speaker was called an *exchange*. Not all the exchanges present in the Map Task Corpus could be used for the rhythmic and temporal analyses reported on in chapters 5 and 6. In this chapter I outline which data was not used, and the reasons why.

Essentially, the problem revolved around uncertainties over whether any given utterance by N could be counted as a response to an utterance by C. Or if it were a response, then could it be treated as a response to the end of the utterance, or to some earlier part of it? These questions were of importance because an analysis of the intervals between speakers' utterances, or between the prominent syllables within those utterances, requires that the correct intervals are being measured. The problem was partly one of corpus size. I isolated over 17,000 exchanges from the corpus, making an assessment of each exchange within the time scale of this research unfeasible. I had to adopt a technique which could be applied to all the data to eliminate broad unwanted categories of that data. I used several methods, based largely on examining a sample set of data, classifying utterances by N as responses to C or not, and using the ratios of positive and negative ases to determine threshold values beyond which an utterance by N was unlikely to be a response to C.

4.2 Units of Measurement

4.2.1 Turns, moves, and utterances

Thus far in this thesis, I have discussed only theoretical issues of conversation and 'turn-taking', and how interlocutors make 'contributions' to a conversation. I used the term *utterance* to approximate to this notion of a contribution, but made no mention of exactly what the definition of an utterance is. This reflected the fact that there is no single definition of a basic unit of analysis, and that very often no exact definition exists. There seems to be an *intuitive* sense of what constitutes an utterance, but a more precise definition is difficult to make, particularly in cases where speakers may be overlapping or interrupting one another. A definition based on the concept of the turn as a unit of discourse has had considerable treatment in the literature, and has been used as the basis of many models of conversation (most notably in the Sacks et al. model of turn-taking). Perhaps one of the more useful units - the TCU - has the drawback that its boundaries only approximate to an intonation unit. The first step here is therefore to define a basic unit of analysis, which will then be used in the data analyses in Chapters 5 and 6.

As I mentioned in Chapter 2, Sinclair & Coulthard (1992) claimed the utterance unit was unnecessary in an analysis of conversation. However, I found that no other unit sufficed for measuring inter-speaker intervals, at least within the context of the Map Task Corpus. There were two other readily-available candidates for a unit of analysis, which taken alone proved inadequate: turn-units and move-units. A turn-unit is at first sight an intuitively appealing concept, because it captures the notion of a contribution by an interlocutor, to which another interlocutor may form a response turn-unit at an appropriate moment. But as I noted in Chapter 2, there are some difficulties with the 'traditional' idea of conversation as a series of turn-units by different interlocutors. A close analysis of real dialogue reveals that participants do not always take it in orderly turns to hold the floor. As Clark (1996) notes, the orderliness of turn-taking is an illusion which arises out of the need for participants to coordinate their joint actions. Furthermore, the placement of turn-unit boundaries in the Map Task Corpus was at best highly subjective, and not reliable

enough for this study: turn boundaries had been set by means of an auditory analysis only, according to the intuitions transcribers had for sequences of contributions from each speaker.

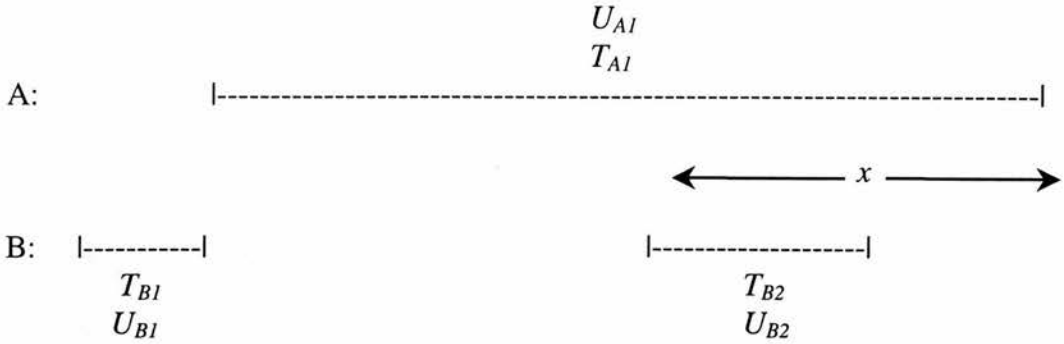
In the corpus, turn-units were based on move-units, which were reliably coded and placed. One might therefore suppose that move-units could be used dircetly to measure inter-speaker intervals. However, taken alone without some concept of speaker switch they were not powerful enough to recognise any interaction between interlocutors.

It is more useful and productive to view conversation as a series of contributions, or utterances, *by each speaker*. Each utterance consists of functionally-defined move units. These contributions may or may not overlap with one another. Moves (as I noted in Chapter 2) have been proposed as the basic free unit of discourse (unfortunately, different authors appear to use different terminology. Sinclair & Coulthard (1992) use a unit called the *act* as their basic unit of analysis, which corresponds to the move as used in the Map Task Corpus). The move may be said to correspond roughly to a syntactic clause, although its boundaries are set according to functional rather than structural considerations. The use of move units has several advantages:

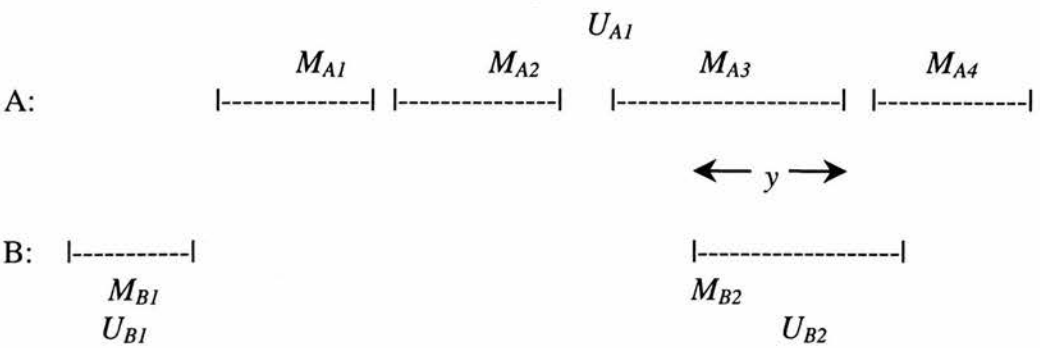
a) Move boundaries were marked to the nearest word boundary, using a combination of auditory and visual (spectrogram and waveform) analyses. This made a move-based analysis of intervals both convenient and accurate.

b) A move-based system allows for greater precision than a turn-based system in measuring inter-speaker intervals where there is overlap between the two speakers. The following representations of the same event demonstrate how different measurement can be made with the two systems, and how the turn-based approach can give a misleading impression of timing.

1a) *Turn-based analysis*



1b) *Move-based analysis*

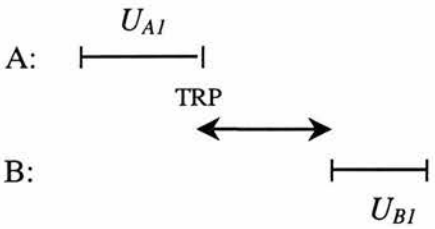


For any given speaker S, U_{Sn} is an utterance, T_{Sn} is a turn, and M_{Sn} is a move, where n is the number of the unit within the example under consideration. In example 1a) U_{A1} is equivalent to T_{A1}. T_{B2} overlaps with U_{A1} by x ms. But in example 1b) T_{A1} has been split up into 4 moves. In that analysis U_{A1} runs from M_{A1} to M_{A3}. The overlap of U_{A1} and U_{B2} is greater in 1a) than in 1b). If the turn-based approach is used, U_{B2} is an early response to the end of U_{A1}.

However, it is possible that U_{B2} is a response to a specific move. By allowing for this possibility the move-based account yields more accurate inter-speaker intervals. But a consequence of this is that there is a problem in deciding which utterance any other utterance might be a response to.

c) Move units appear to fit in well with theoretical accounts of speaker-switching. Interlocutors use turn constructional units (TCU's)[16] in the coordination of their utterances (Ford & Thompson, 1995). TCU's are bounded by transition relevance places (TRPs) (see Chapter 2 for a fuller theoretical discussion of this). Responses are thought to be made to TRPs. It therefore seems reasonable to assume that any measurements of intervals between speakers' utterances should be taken between A's TRP, and the start of speaker U_{B1}, as shown in 2):

2)



Although no precise definition exists of TRPs, their placement appears to depend on a complex interaction of several factors, such as intonation, syntax, pragmatics, gesture and gaze. But a reasonable claim could be made that they coincide roughly with move boundaries, and vice versa (see Chapter 2). However, this relationship is not a clear one, because move units are defined in terms of their function, and make no appeal to gaze or intonation. Certainly, turn boundaries may also coincide with TRPs, but it is not necessarily true that all TRPs must coincide even approximately with turn boundaries. Turn boundaries may reflect an attempt at capturing *actual* speaker switches, but they do not represent points of *potential* speaker switch.

I decided therefore to use an utterance as a basic unit of analysis, where an utterance consisted of a series of move-units by one speaker only.

---

[16]Unfortunately the name TCU presupposes that a turn unit is a genuine unit of conversation. The current analysis moves away from a reliance on turn units, but the expresison TCU is still retained because it is used extensively in the literature.

4.2.2 Definition of an Utterance

Utterances were defined in terms of move sequencing across speakers. Intervals were calculated only whenever two successive moves were uttered by different speakers. The definition of an utterance as used here is therefore as follows:

Definition 1

An utterance is a sequence of one or more moves by speaker B, where the start points of the first move (M_{BI}) and last move (M_{Bj}) in B's utterance (U_B) lie between the start points of two of speaker A's moves (M_{AI} and M_{AI} respectively), except at the beginning or end of a conversation.

Note that this definition allows for temporal overlap: M_{BI} starts after M_{AI} begins, but not necessarily after M_{AI} ends. M_{Bj} may continue while A is uttering U_A. However, inter-speaker intervals lie between the offset of one speaker's utterance, and the onset of the other's. An *exchange* is a pair of utterances surrounding an inter-speaker interval. This is shown in 3) below, where a * represents the start point of each move:

3)

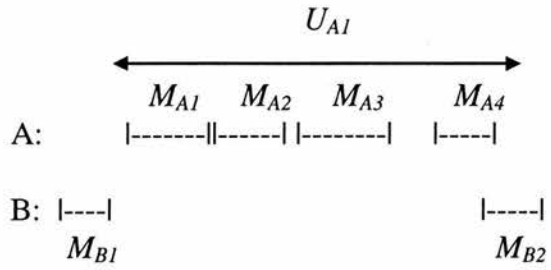
```
        1    2    3    4
A:      *    *    *    *

B:  *                     *
    1                     2
```

For the purposes of calculating the boundaries of an utterance it therefore only matters where the *start points* of each speaker's move are located relative to one another. The corpus had previously been labelled with move boundaries, and since each move includes time-stamped words it was possible to isolate the start and end times of each move for each speaker in the corpus. The start and end points of a move were assumed to coincide with the beginning and end of *speech*. The corpus was marked for lipsmacks and intakes/out-takes of breath, which often appeared to indicate the start of a contribution (or the intention to start a contribution).

Nevertheless, it was felt that these noises could not be relied upon because they did not *necessarily* indicate the start of an utterance.

All the moves were separated into utterances according to the rule above. Consider example 4) below.

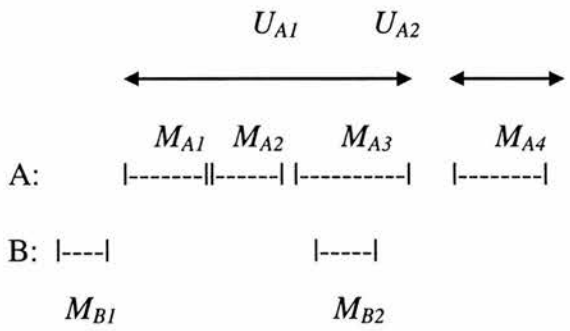
4)



In example 4) above the sequence of moves numbered 1-4 by speaker A comprise an utterance, U_{A1}, because the start points of M_{B1} and M_{B2} bound the start points of M_{A1} and M_{A4}, which are the first and last moves in the sequence. Note that this means that M_{B2} overlaps with part of M_{A4}, but does not break the sequence of moves by A. M_{B2} would therefore count as a case of overlap with U_{A1}, where the overlap is measured from the start of M_{B2} to the end of M_{A4}.

However in example 5) below, only M_{A1} to M_{A3} comprise U_{A1}, because now one of B's moves (M_{B2}) starts before the onset of M_{A4}. M_{A4} would therefore be counted as a separate utterance (U_{A2}).

5)



In many respects, therefore, an utterance as defined here is similar to a turn-unit. The crucial difference lies in the fact that an utterance boundary is defined purely in terms of speaker-switching, calculated to the nearest move boundary. A turn-unit may be defined in terms of speaker-switching, but need not be. They may be defined in terms of subjective criteria such as when a speaker has finished saying what he or she wishes to say, before possibly starting speaking again.

4.2.3 Definition of 'response'

A basic concern was whether an utterance by speaker B (U_B) could be counted as a response to an utterance by speaker A (U_A). And if U_B were a response, could it be treated as a response to the end of the utterance, or to some earlier part of it? The definition of an utterance has its difficulties. The definition of a *response* to an utterance is equally problematic. Generally, we can say that:

Definition 2

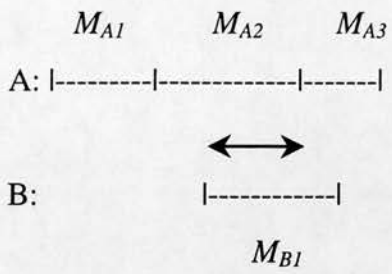
An utterance U_B is a response to an utterance U_A when U_B is in some way elicited by the content of U_A. It is not enough for U_B to have occurred at some point after the start of U_A for it to be a response to U_A.

Of course, the problem lies in knowing when U_B has been elicited by U_A and when it has not. The only reliable method within the constraints of this analysis was to use move coding, which gave at least a basic form of semantic representation.

4.2.4 Uncertain Responses

An utterance U_B is a response to an utterance U_A when U_B is elicited by the content of U_A. The difficulties in determining this arose not only from the problem of how to be certain that U_B was elicited by the content of U_A, but also from the way that this analysis split chains of moves into utterances. According to the definition used in this analysis, in 6) below M_{A1} and M_{A2} would be treated as an utterance, and M_{A3} would be treated separately.

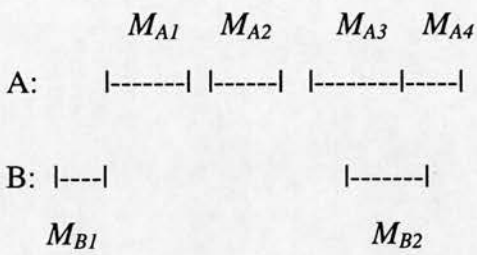
6)



But it is not at all certain whether this treatment is correct, for two reasons:

a) Under the general rule some moves would be treated as separate utterances in their own right, even when it appeared that they should not to be, as shown in example 7a) below.

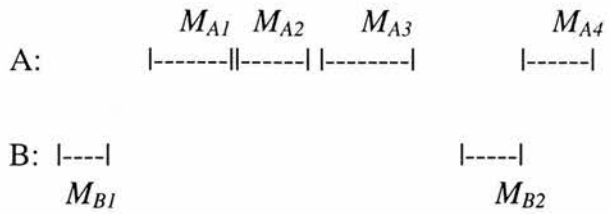
7a)



In 7a), moves M_{A1}, M_{A2}, and M_{A3} are treated as one complete utterance, as are M_{A4}, M_{B1}, and M_{B2}. Since there is no pause between M_{A3} and M_{A4}, it is difficult to claim that M_{A4} should be treated as a separate utterance. Or at least it seems harder to claim this than if, for example, the interval between M_{A3} and M_{A4} were considerably larger.

The interval between M_{A3} and M_{A4} is not the only limiting factor, however. Two moves by the same speaker may still be counted as falling in the same utterance even when there is a large gap between them, if the second move starts at the same time - or nearly so - as a move by another speaker:

7b)



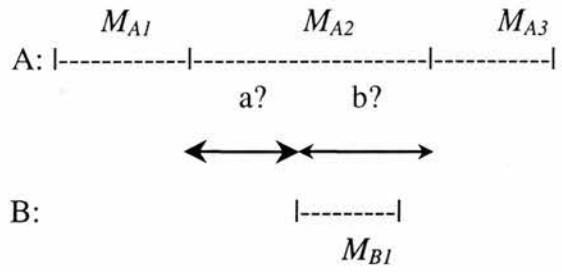
In 7b) we see that although there is a large gap between M_{A3} and M_{A4}, M_{A4} starts only slightly after M_{B2}. It is unlikely that M_{A4} could be a response to M_{B2} if the difference between their start times were small.

In effect, the question arises in both 7a) and 7b) as to whether M_{A4} is a response to M_{B2} (and therefore a valid utterance in its own right), or whether it was not likely to be a direct response to anything in itself - simply a continuation of the utterance starting with M_{A1}.

If the general utterance boundary rule is overridden, and M_{A4} is counted as just a continuation of M_{A1} to M_{A3}, the interval would *still* be measured between the end of M_{A3} and the start of M_{B2}. The difference would result in the loss of M_{A4} as an utterance, and therefore the loss of the interval between the end of M_{B2} and the start of M_{A4}.

b) Even if we conclude that a move is definitely a response to the other speaker's utterance, there still remains the question of where in that utterance the interval should be measured from. As is shown in 8) below, the interval between U_A and U_B could be measured from the end of M_{A1} to the start of M_{B1} (giving a positive interval) or it could be measured from the end of M_{A2} to the start of M_{B1} (so giving overlap). Both treatments are equally plausible, depending on whether M_{B1} is a response to M_{A1} or to M_{A2}. The problem here then is to decide for any given move which other move it is responding to.

8)



This problem still exists with a turn-based approach. It is simply that it is not immediately apparent because there is only one point in the turn (its end) from which measurements could be taken. A turn-based approach therefore not only masks these problems, but as mentioned earlier it yields potentially less accurate inter-speaker intervals.

4.3 Data Reduction

4.3.1 Elimination of erroneous data

I calculated inter-speaker intervals from the end of U_A to the start of U_B. This gave intervals for 18248 exchanges from the whole corpus. The first step in reducing the data was to eliminate exchanges containing errors or misclassifications, as follows:

a) Because of incomplete move coding in the corpus, 896 move categories were classified as either 'uncoded' or 'unclassifiable'. Any exchange containing such a move was excluded.

b) In 273 cases the move was not recognised by the coder because the only word or words in the move belonged to the category 'unknown'. If a move was not properly recognised, the exchange in which it occurred was not used.

This left 17079 valid exchanges. Next, I had to decide which utterance any other utterance was a response to, and whether some utterances could be counted as

responses to anything at all (for example, whether an utterance following a backchannel signal could be treated as a response to it). The general definition of an utterance therefore was modified according to several further criteria, as outlined below.

4.3.2 Criteria for Response Determination

The solution to the two problems noted in section 4.2.4 above was to take a sample of exchanges, and to determine for each one whether the moves concerned were a) responses to other moves, and b) if so, which moves it might reasonably be thought that they were responding to. The objective was to use a set of criteria to establish patterns of inter-speaker interval durations within a sample data set, which could be used to reduce the entire data set.

The response/non-response decision task was achieved using two basic methods. Subjective decisions were made on the basis of three criteria: semantic content of the exchange, and the move and game coding in each exchange in the sample. Some decisions were made automatically, according to specific patterns of move and feature coding in each exchange. This automatic filtering acted irrespective of possible subjectively-based response/non-response decisions, and therefore overrode the subjective decision making process. The sections below list in more detail the criteria used in the filtering process.

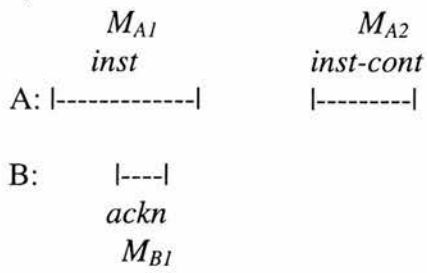
4.3.2.1 Automatic reduction

Two classes of apparent exchanges were removed automatically:

a) *The 'cont' Feature*

Some apparent utterances in the data were 'linked' to one another because the moves contained in the linked utterances were each a part of the same overall move, and yet were coded separately. Such moves were marked in the Map Task Corpus with the 'cont' feature. In effect, the 'cont' feature connects two moves that should logically be one (Kowtko, 1997).

9)



In example 9) M_{A2} is marked with the 'cont' feature, and is treated as a continuation of M_{A1}. It therefore is not taken to be a separate utterance, despite the general definition of an utterance used here. Any exchange which contained an utterance which was linked to another utterance through the 'cont' feature was eliminated from the analysis.

Although the 'cont' feature effectively acts as a label preventing a move being counted as a separate utterance in itself, it cannot be said that the absence of this feature means that a move must be a separate utterance because it would be quite possible for speaker A to start a new move, and yet not respond to speaker B. The absence of a 'cont' feature simply means that there is the potential for a move to be treated as a separate utterance.

b) *Backchannelling*

The back-channel is essentially a form of communication used by participants as a social tool to demonstrate a sense of cohesiveness and attentiveness. Unlike the main-channel of communication, the back-channel has no goal-oriented informational content, and does not pose a challenge for the conversational floor. The distinction between main- and back-channels, however, is not necessarily clear because many utterances such as '*yeah*' or '*mm-hmm*' may act as confirmations, and therefore to some extent like a full contribution in the main-channel, as in 10):

10) (*Figures in brackets denote time in seconds between utterances*)

G: And then the carpenter's cottage is to the left of a ravine

(0.906)

F: mm-hmm

(0.133)

G: I want you to take the line underneath the carpenter's cottage

Treatment of the backchannel is therefore problematic. Should the backchannel be treated alongside the main channel of communication in a consideration of the response of one move to another? Although the backchannel is non-competitive, it nevertheless provides C with positive feedback that his or her signal has been received and understood (recall the discussion of Clark's theory of conversation in Chapter 2 of this thesis).

Further, can utterances be responses to utterances ending in, or consisting entirely of, backchannel signals? Take the following example:

11)

	U_{A1} *instruct*	U_{A2} *instruct*
A:	starting off we are above a caravan park	we are going to...
B:	mmhmm	
	U_{B1} *acknowledge*	

Here, A gives an instruction (U_{A1}), B then 'replies' with an acknowledgement (U_{B1}), which is a backchannelled utterance. Note that the definition of backchannelling here is based on the Map Task move coding. *Acknowledge* moves have been treated as backchannels (Kowtko, personal communication). In example 11) speaker A 'replies' to the backchannelling with another instruction (U_{A2}). It is unclear whether U_{A2} is a true response to U_{B1}. If it were a response, it would appear to be very much a different type of response from, for example, a response to a *query-yn*. But a speaker

may be said to be responding to a backchannel signal inasmuch as communication would break down if that response were not present.

The 'repo' feature was also used in the definition of backchannelling. A move with this feature is a move which repeats part or all of the other speaker's previous utterance. When on a *check* or *acknowledge* it functions to check accurate transmission. e.g. 12) below. By definition moves with this feature have to be responses to a previous utterance.

12)

(*other speaker's previous utterance is in angled brackets*)

<Ehm, right and you're turning left up there.>

"Turning left?'' CHECK--repo

(Kowtko, 1997)

Backchannel signals were included with the rest of the data in this analysis only when they came second in an exchange, as in U_B in example 12) above. Utterances which followed utterances ending in, or consisting entirely of, backchannel signals were not treated as responses to them. This was because by Defintion 2 earlier a response must be elicited by the content of an utterance. Such exchanges were therefore omitted from the main analysis. That is, the hypothesis here is that backchannelled utterances are not simply randomly placed contributions, but are timed to coincide as near as possible with relevant TRPs in C's utterance.

4.3.2.2 Reduction criteria set by examining samples

I isolated a sample of 441 exchanges from the corpus for manual inspection. The intention was to determine exchanges where an utterance U_B might reasonably be considered a response to an utterance U_A, and exchanges where it might not. This task required some degree of subjective decision-making, although I also applied the automatic criteria outlined above to the sample data set.

It was possible to reduce the subjective decision-making process to include three broad considerations: move coding, the presence of a game boundary, and the semantics/ pragmatics present in each exchange.

a) *Move Coding*

The move coding of each move in an utterance gave some clue as to whether U_B might be a response to U_A. For example, an utterance starting with a *reply-y* move is more likely to follow and respond to an utterance ending with a *query-yn* move, than to follow and respond to an utterance ending with an *instruct* move. In particular, *acknowledge* moves tend to respond to *instruct* moves, largely because of the goal-oriented nature of the Map Task dialogues.

An analysis of the distribution of move categories revealed the following as some of the more common exchange pairs, where the figures represent the percentage distribution of each category out of the 17, 079 exchanges used in the analysis:

acknowledge/ align	3%	*instruct/ acknowledge*	10.7%
acknowledge/ instruct	7.6%	*reply-y/ acknowledge*	3.7%
check/ reply-y	6.1%	*reply-y/ instruct*	2.4%
clarify/ acknowledge	3.1%	*reply-w/ acknowledge*	2.3%
explain/ acknowledge	4.2%	*query-yn/ reply-n*	2.5%

There are 132 exchange pair types (12 move classes, giving 144 pairs, less 12 because of the condition eliminating moves following backchannelled - or *acknowledge* - moves). The 10 pairs here represent 45.6% of the exchanges used in the overall analysis.

The move coding allowed for an analysis of two basic types of move - *initiator* moves and *response* moves (see Chapter 3 for a fuller discussion of this). These two classes could therefore be used to determine whether any utterance was likely to be a response or not, since utterances starting with *initiator*-type moves are likely to start a new sub-task in the conversation, and therefore are less likely to be a

response to a previous utterance. However, *initiator* and *response* coding could not be used accurately as blanket determiners of response because the inclusion of a move category into one of the two classes was only general, and specific cases may have differed. For example:

13)

G: Have you got a haystack on your map?

F: Yeah

G: Right just move straight down from there, then

F: Past the blacksmith?

F's utterance "*Past the blacksmith?*" is a *query-yn* move, which falls in the initiator class of moves. But in this example, it appears to act as a response to G's previous utterance. It asks for a confirmation of where "*straight down from there*" is.

b) *Semantics/ Pragmatics*

Determining whether utterance U_B is a response to U_A requires consideration of semantic and pragmatic aspects of both utterances. Unfortunately, the necessary semantic and pragmatic criteria are difficult to define. They can only be accounted for in any explicit way in the Map Task Corpus coding scheme via move categories. But this is not enough in itself, as example 13) above shows, because the discourse function that an utterance may have need not be related to its status as a response. In this analysis I was forced to rely on a sense of 'pragmatic relatedness' between two utterances. In 13) the relatedness is concerned with F's asking for confirmation that F has followed the instructions correctly. A similar example is:

14)

F: I'm in between the remote village and the pyramid

G: Are you?

Again, G's utterance is a *query-yn* (that is, a prototypical *initiator*-type move). Yet it has the status of a response to F's instruction. Clearly, certain types of move can act simultaneously as responses (in that they may answer a question or reply to an instruction) and initiators (in that they require a response themselves). Clark's theory of conversation predicts that utterances generally (with the exception of feedback signals) act to signal a response to other speakers' utterances *and* require some feedback in return from the other participant.

c) *Game Boundaries*

Game boundaries were also used to determine whether U_B was a response to a U_A. The first utterance in any game was the least likely to be a direct response to any previous utterance. Nevertheless, some game-initial utterances appeared to be acting as a response to a previous utterance. Example 15) shows how an utterance (consisting entirely of a *check* move) which begins a new game also bears a 'repo' feature. As was noted earlier the 'repo' feature indicates that an utterance is definitely a response to an utterance by the other speaker.

15)

[Follower. *check repo* STARTGAME]

F: p... slightly northeast

In 15) it would be hard to justify the claim that the utterance was not responding to anything, therefore. The presence of a game-initial utterance was therefore only used as a guideline, which could be qualified by the move type and feature type of the first move in that utterance.

4.3.2.3 Setting of Cut-off Points Based on Sample Data

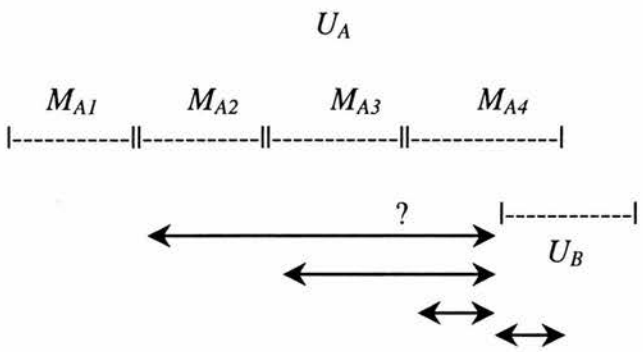
I used the 441 sample exchanges taken from the corpus for a manual-based analysis. The results were used to generate temporal cut-off points which could be applied to the whole corpus of 17079 exchanges, simultaneously with the automatic criteria. The sample data was split into three categories, according to the type of temporal

information considered. These were *excessive overlap*, *nearly simultaneous onsets*, and *uncoded continuations*.

a) *Excessive overlap.*

The premise here was that U_B is less likely to be a response to U_A if it starts well before U_A ends. Often, U_B may be considered to be a response to *some part of* U_A, but not necessarily to the last move. When there are several moves in an utterance, there is no reason why the second utterance could not be a response to any of those moves, as 16) shows.

16)



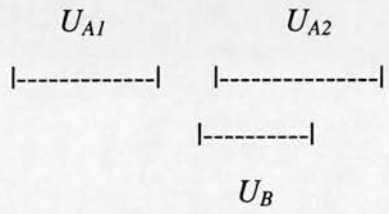
Unfortunately, it proved impossible in this study to say with any certainty which part of U_A that U_B was a response to, because this would have required a far more subtle analysis of the semantics and pragmatics of the conversations than was readily available. The most that could be done was to attempt to determine whether U_B would be a response to the final move in U_A.

b) *Nearly simultaneous onsets*

If U_A and U_B start within only a few hundred milliseconds of one another, the delay may be insufficient for B to have interpreted U_A and to have prepared a response. Near-simultaneous onsets may occur accidentally, usually when speaker B makes some response to an utterance by A just before A continues on from his first move. Or they may be used deliberately to 'shadow' another speaker's utterance, and to

reinforce its impact. This latter case proves to be problematic, because instances of cooperative shadowing can only be distinguished from accidentally simultaneous moves by their respective content. That is, if both utterances share the same or similar words, then it is likely that one utterance is a deliberate shadowing of the other.

17)

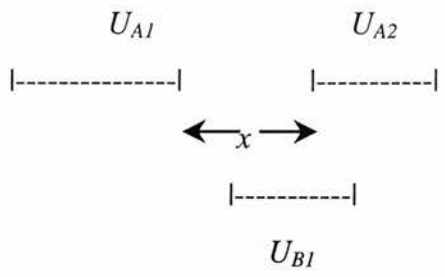


In 17) the interval between the start of U_B and the start of U_{A2} is presumably too small for A to have reacted to the start of U_B. U_{A2} is therefore either not a separate utterance, but part of the same utterance as U_{A1}, or begins a new topic of conversation.

c) *Uncoded continuations*

If an utterance (U_{A2}) starts very soon after another utterance (U_{A1}) by the same speaker, the second utterance may in fact be a continuation of the first, and not a separate utterance at all. In effect, this category captured those exchanges which ought to have been coded as continuations in the corpus, but were not. Example 18) below shows the *inter-utterance interval*, *x*. Note that the position of U_{B1} does not figure in the calculation of inter-move interval. We are claiming that when the interval between U_{A1} and U_{A2} is above a certain threshold, each is a part of a different utterance, and that U_{A2} may be a response to U_{B1}. When the interval between U_{A1} and U_{A2} is below the threshold, they are in fact just part of the same utterance. The interval between the end of U_B and the start of U_{A2} is therefore no longer relevant for the calculation of inter-speaker interval.

18)



4.3.3 Method and Results for the Analysis of Sample Data

4.3.3.1 Excessive Overlap

Exchanges were selected for the sample data where the inter-speaker interval fell within 1000ms intervals between –5000ms and 0ms. The ranges were separated into these groups purely for convenience, although they could equally have been grouped together as all exchanges involving negative intervals. The exchanges falling within each group were listed in order of quad, eye contact, conversation number, and move number (each exchange consisted of two moves - one by each speaker. Both move numbers were listed, although ordering was done according to the first move number in each exchange). I selected every 20th exchange in the list for further analysis (because this gave a sufficiently large sample, without being too large), giving a total of 236 exchanges in the sample set. The dialogue corresponding to each exchange number was isolated from the appropriate SGML file, and I applied both the subjective and automatic criteria outlined earlier.

Two examples of exchanges follow, together with a rationale of the decision made in each case. The representations in the examples that follow show the utterances of each speaker in one column. Each word is listed separately on one line, together with its start time, in seconds, relative to the whole conversation. The first line of each utterance indicates: which participant was the speaker, the move number, and the move category (plus any move features). The two columns are approximately aligned to give the impression of the intervals between utterances.

19)

giver move 26 *instruct*

START=82.8227 **and**

START=82.9315 **then**

START=83.1423 **we're**

START=83.3136 **going**

START=83.6001 **to**

START=83.7615 **turn**

START=84.3149 PAUSE[0.7937]

START=85.1086 **to**

START=85.2644 **the**

START=85.3541 **west**

START=86.0606 PAUSE[0.4315]

START=86.4921 **on**

START=86.6107 **a**

START=86.6864 **curvature**

START=87.5297 PAUSE[0.8629]

follower move 27 *acknowledge*

START=86.4542 **okay**

START=88.3926 **right**

START=88.7351 **sort of**

In example 19), the second move (no. 27) is by the follower. Its start time is 86.4542 seconds into the task. This can be aligned to the nearest word in the giver's move (no.26), which is after the word "west" (which finishes at 86.0606 seconds), but before "on" (start = 86.4921). The fact that the giver here pauses for 0.4315 seconds, and that it is in this pause that the follower starts her move, seems to indicate that move 27 may be a response to the *first part* of move 26 - "and then we're going to turn to the west", rather than an overlap in anticipation of the end of the whole move.

Move 27 is most likely a response to at least a part of move 26 because a) this exchange pair is an *instruct/ acknowledge* type, b) move 27 does not start a new game. However, because move 26 does not respond to the end of move 27, this exchange would not be considered further. The overlap between move 26 and move 27 is accidental, but nevertheless represents a genuine attempt by the follower to

make an utterance at a strategic point. In this sense, the second 'chunk' of move 26 (26a, starting after the word '*west*') is a real unit. However, to assess the real interval between 26a and 27 would require a re-working of the entire corpus - something which must be reserved for future research.

20)

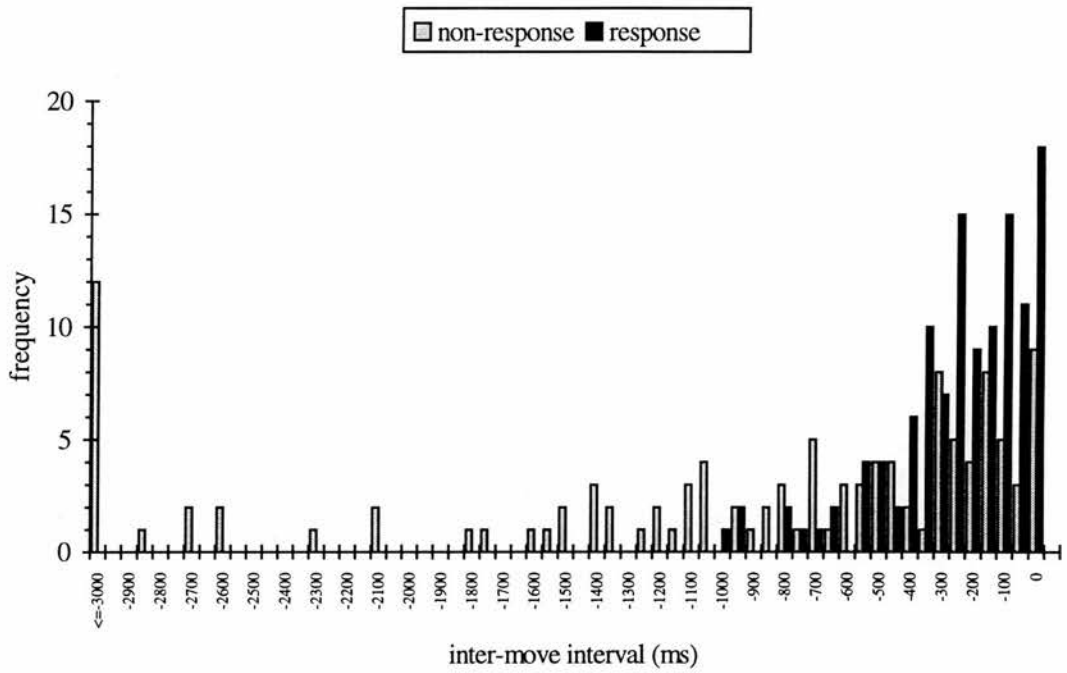
giver move 52 *instruct cont*
START=75.8580 **missing**
START=76.2780 PAUSE[0.1847]
START=76.4134 **is**
move 54 *query-yn* STARTGAME
START=76.5842 **is**
START=76.6279 PAUSE[0.3517]
START=76.9796 **that**
START=77.0992 **where**
START=77.1991 **your**
START=77.2497 PAUSE[0.0234]
START=77.2731 **carpenter's**
START=77.6805 PAUSE[0.0184]
..... [lines omitted to save space]
START=79.4594 **the**
START=79.4967 **burnt**
START=79.6827 **forest**

follower move 53 *acknowledge*
START=76.6676 **right**

In example 20) above, the interval under consideration is between moves 54 and 53, in that order. However, move 53 starts only 83 ms after move 54 starts (after the word "is" in 53). It seems highly unlikely that move 53 is a response to the whole of move 54, with some 3480.7ms of overlap. Rather, it is a response to move 52. That this is so can be supported by noting the move types concerned. Move 52 is an *instruct*, move 54 is a *query-yn*, and move 53 is an *acknowledge*. One might expect a *query-yn* to be followed by a *reply-y* or *reply-n*, not an *acknowledge* move. But one

could expect an *acknowledge* move to follow an *instruct* move, as is the case if move 53 is considered a response to move 52. Therefore, the decision here would be to disregard the 3480.7ms overlap between moves 54 and 53.

Results from the Determination of Response in Overlap Cases



no response: n = 116 mean = -1222ms, SD = 1803ms
response: n = 120 mean = -271ms, SD = 225ms

Figure 1 - Histogram of the distribution of inter-speaker intervals. The upper limits of each bin are labelled.

Figure 1 shows that non-responses have a broad general distribution: the presence of a non-response is equally likely no matter what the inter-speaker interval is. The distribution of non-responses can be ignored, because there were no definite responses when the interval was less than -1000ms. All exchanges with intervals below this -1000ms cut-off point were eliminated from the analysis. Table 1 below shows the frequencies of non-response/response observations above and below -1000ms, and that there is a significant difference between them. Of all cases passed by this filter, the ratio of responses to non-responses was 1.62. A total of 424 exchanges with overlap greater than 1000ms were excluded from the corpus,

although a further 439 were excluded in common with either or both of the other filters.

Table 1. response/non-responses by interval length

	non-response	response	total
<-1000ms	42	0	42
≥-1000ms	74	120	194
total	116	120	236

$\chi^2 = 50.42$, df = 1, $p << 0.001$

4.3.3.2 Near simultaneous onsets

I isolated 6252 exchanges from the Map Task Corpus, where the start of U_B occurred less than or equal to 1000ms before the start of U_{A2}. These were ordered according to their respective quad numbers, eye contact, conversation number, and move number. Every 50th exchange was taken for further analysis, giving a sample of 126 exchanges. The dialogue corresponding to each exchange was extracted from the SGML files. For each of these, a decision was made as to whether U_{A2} could reasonably be counted as a response to U_B, or instead as a continuation of U_{A1}.

In 21) below, the exchange in question here involves moves 325 and 326. Move 324 is a *query-yn*, 325 is a *reply-n*, and 326 an *acknowledge*. It seems plausible that 326 is a response to the *reply-n* move in 325, and acts to reinforce the reply to the original question in 324. Also, the word '*no*' in 326 occurs after the end of move 325 - indicating that 326 may be a valid response. This data combined would lead to the decision that 326 is a valid response to 325.

21)

follower move 324 *query-yn*
START=588.7283 **have**
START=588.8090 **you**
START=588.8897 **got**
START=589.0000 **whitewashed**
START=589.4134 **cottage**
589.7543

giver move 325 *reply-n*
START=590.3282 **no**
590.6291

follower move 326 *acknowledge*
START=590.9187 **no**

In 22) below, moves 91 and 90 occur almost simultaneously (move 90 starts 173ms before move 91). The problem is therefore to decide whether 91 could feasibly be regarded as a response to 90. There are two reasons for thinking that it could be.

22)

giver move 89 *acknowledge*
START=211.9829 **mmhmm**
212.2365

follower move 90 *query-yn*
STARTGAME
START=212.1739 **is**
START=212.2742 **that**
START=212.3843 **us**
START=212.5323 **finished**

giver move 91 *clarify*
START=212.3471 **take**
START=212.5362 **it**
START=212.6548 **to**
START=212.7536 **the**
START=212.8359 **lighthouse**

First, move 90 starts a new game. Second, the exchange pair is *query-yn/ clarify* type. Although this is not one of the most common exchange pairs, it is nevertheless quite

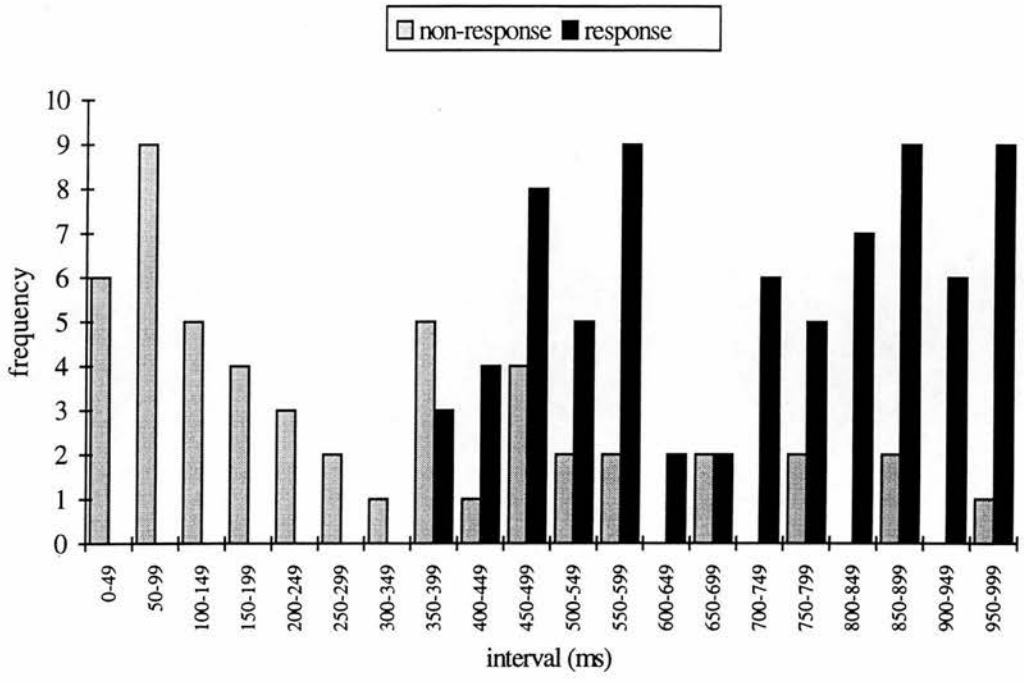
conceivable that a *clarify* move could act as a response to a *query-yn* move, in instances where a question makes a wrong assumption, for example.

However, an analysis of the semantics and pragmatics seem to show that move 91 is unlikely to be a response to move 90. It is difficult, in the context of the Map Task, to see how an utterance '*take it to the lighthouse*' could be a valid response to the utterance '*is that us finished?*'. So, the exchange in 22) would be disregarded on semantic and pragmatic grounds, even when other evidence points to the contrary.

Example 23) below shows the nearly-simultaneous moves 150 and 151. 150 starts 142ms before move 151. Two points become immediately apparent. The exchange pair is a *reply-y/ query-yn* which is not particularly common, and does not appear to be typical of an initiator-response pattern. Also, move 151 starts a new game. It is possible that move 151 is a rapid response to 150, and it could act as a request for more information. However, on balance it seems more likely that 151 simply continues, or elaborates on the follower's previous move, and does not act as a response to move 150.

23)

	giver move 150 *reply-y*
follower move 151 *query-yn*	START=290.7101 **mmhmm**
STARTGAME	291.0342
START=290.8523 **under**	giver move 152 *instruct*
START=291.1322 **it**	STARTGAME
291.3101	START=291.2786 **go**
	START=291.5080 **up**



non-response: n = 51, mean = 311ms, SD = 259ms
response: n = 75, mean = 710ms, SD = 192ms

Figure 2 - The distribution of intervals between 'simultaneous' moves

Figure 2 shows that there are only non-responses when the start times of two utterances differ by less than 350ms. Above this level, there are primarily responses, although there are still some non-responses. It seemed that when the difference between the start points of two utterances was less than 350ms, the utterance that started later would not be a response to the first. A cut-off point of 350ms was set, and any exchanges where the moves had a difference in start-points less than this were eliminated from the analysis. Table 2 below shows the frequencies of non-response/response observations above and below 350ms, and that there is a significant difference between them.

Table 2. response/non-responses by interval length

	non-response	response	total
<350ms	30	0	30
≥350ms	21	75	96
total	51	75	126

$\chi^2 = 54.72$, df = 1, $p << 0.001$

In the cases passed by the filter, the ratio of responses to non-responses was 3.57. In total, 567 exchanges, where the onsets of the utterances were less than 350ms, were excluded from the corpus, although a further 798 were excluded in common with either or both of the other filters.

4.3.3.3 Uncoded continuations

I isolated each neighbouring pair of exchanges in the Map Task Corpus, and calculated the inter-move interval. This gave 16975 inter-move intervals. I used only those cases where the inter-move interval was greater than or equal to 0ms, and less than or equal to 1000ms.

There were in total 3912 exchange pairs where the inter-move interval was between 0ms, and 1000ms. These were ordered according to their respective quad numbers, ±Eyecontact, conversation number, and move number. Once ordered, every 50th case was selected, giving a sample of 79 exchange pairs, which were analysed using both subjective criteria, and the automatic filters.

As an example of the sort of decisions that were made in determining response, consider 24):

24)

giver move 80 *query-yn*

START=96.7588 **and**

START=97.1265 <unknown>

START=97.3064 **the**

START=97.4136 **lagoon**

START=97.7453 **on**

START=97.8540 **the**

START=97.9097 **other**

START=98.0248 **side**

START=98.3499 **to**

START=98.4254 **the**

START=98.4915 **lagoon**

START=98.8481 **right**

99.0108

follower move 81 *reply-y*

START=98.1897 **yes**

98.7413

giver move 83 *ready* STARTGAME

START=99.4405 **well**

In 24) it is the interval between the two moves by the giver (moves 80 and 83) that is under consideration. It should be noted that the follower's move (move 81 - the word "yes") starts at 98.1897 seconds, which is during the word "side" in move 80, and finishes at 98.7413 seconds (during the word "lagoon" by the other speaker). Move 81 therefore occurs at the same time as move 80. In move 83, the giver uses a *ready* move (the utterance "well"). They generally introduce new games, as in this example. Because of the *ready* move, and the start of a new game, move 83 in example 24) would not be treated as a response to move 80.

In example 25) below the inter-move interval between moves 150 and 152 is under examination. The giver says '*mmhmm*', presumably in reply to the follower's

question. The follower then says '*under it*' (which may or may not be a reply to move 150). The point here is whether the giver's 'go up' is a response to '*under it*'. There are at least two reasons for thinking that it is not, and that that exchange should be eliminated from the analysis. The first reason is that move 152 starts a new game. Also, the exchange is a *query-yn/instruct* type, which is not common, and does not appear to form a likely initiator-response pair. More important, in move 153 the giver appears to cut short an instruction, and continue without pausing with a *reply-n* - '*no*'. It seems likely that it is *this* move, and not move 152, which is the response to move 151.

25)

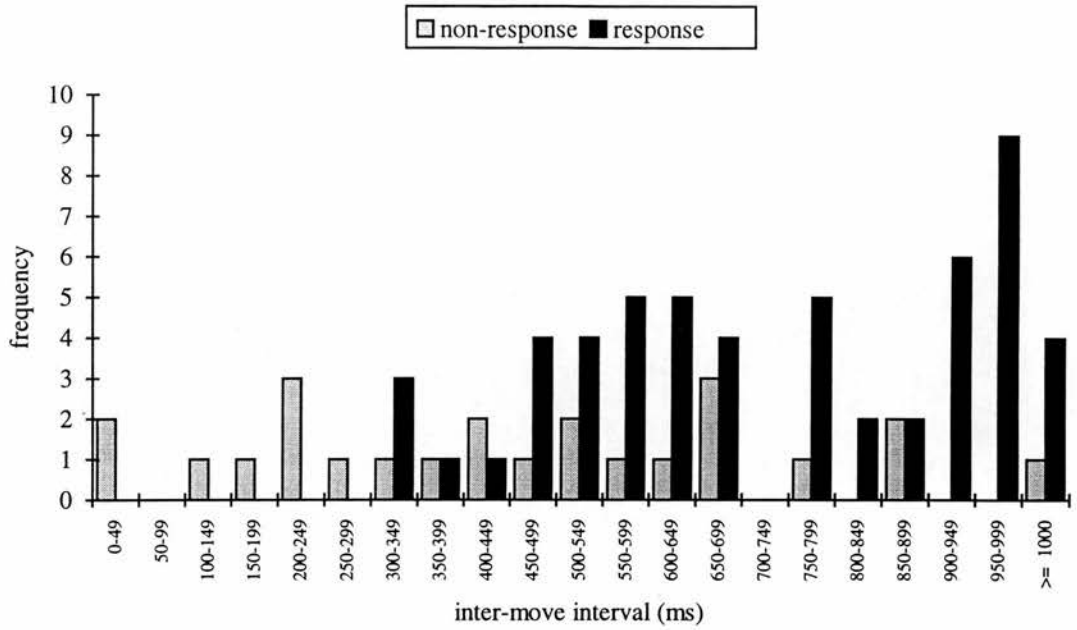
follower move 149 *query-w*
START=289.7752 **of**
START=289.9069 **that**
START=290.1474 **or**
START=290.2709 **what**
290.5558

giver move 150 *reply-y*
START=290.7101 **mmhmm**
291.0342

follower move 151 *query-yn*
STARTGAME
START=290.8523 **under**
START=291.1322**it**

giver move 152 *instruct*
STARTGAME
START=291.2786 **go**
START=291.5080 **up**

giver move 153 *reply-n*
START=291.7091 **no**



non-response: n = 24, mean = 419ms, SD = 272ms
response: n = 55, mean = 681ms, SD = 217ms

Figure 3 - Frequencies of inter-move interval length when responses and non-responses followed the interval (n = 79)

Figure 3 shows that the non-response cases are few throughout the sample range. Inter-move interval therefore does not appear to be related to the chances of finding non-response utterances. However, response cases do not occur below an inter-move interval of 300ms. I therefore used this cut-off point. Below 300ms, I ignored the interval: the two moves by the same speaker are sufficiently close to one another that the second is unlikely to be a response to an utterance by the other speaker, so that the two moves were considered as separate utterances. But above 300ms an utterance may or may not count as a response. Table 3 below shows the frequencies of non-response/response observations above and below 300ms, and that there is a significant difference between them.

Table 3. response/non-responses by interval length

	non-response	response	total
<300ms	7	0	7
≥300ms	16	55	71
total	23	55	78

$\chi^2 = 15.55$, df = 1, $p < 0.001$

Of the cases passed by this filter, the ratio of responses to non-responses was 3.44. In total, 238 exchanges, where the intervals between the utterances were less than 300ms, were excluded from the corpus, although a further 493 were excluded in common with either or both of the other filters.

4.3.4 Validation Study

4.3.4.1 Method and Materials

Four judges were presented with a randomly-selected sample (n = 60) of the cases originally examined, and a randomly-selected sample of cases not previously judged (n = 60). The judges could both read transcripts of the utterances on a computer screen, and hear all or part of the utterances as many times as they liked by clicking a cursor on the relevant text, before making a decision. As well as being presented with the target utterances (marked in red on the screen), the judges were also presented with several utterances which preceded and followed the target in the dialogue (marked in blue).

Of the 120 stimuli cases, 40 met the requirements of the excessive overlap and near-simultaneous onset filters described earlier. 40 did not meet the requirements of the excessive overlap filter (filter 1) alone, and another 40 did not meet the requirements of the near-simultaneous onset filter (filter 2) alone. I decided not to use cases accepted or rejected on the basis of the uncoded continuations filter (filter 3) for two reasons. First, the total number of cases in the judgement task had to

be kept reasonably low because of time-constraints on the judges. Second, I decided that for purposes of validation, cases captured by filter 3 were largely captured also by the other two filters (58% of all cases were captured by one or both of filters 1 and 2 as well as filter 3). 45% of cases passed by filter 3 would have been rejected by the automatic filters, as opposed to only 12% of the cases passed by filter 1.

Each judge was instructed to decide whether an utterance U_B was a response to the final part (or move) in a previous utterance, U_A, by another speaker. They were told to use their subjective judgement as much as possible, and no rigid selection criteria were given on which they could base their judgements.

4.3.4.2 Results

There was generally good agreement between the four judges. Table 4 shows values of kappa which hold between judges a, b, c, and d, using the 60 stimuli not used in the original data reduction study. The values of kappa here were generally good (between K = .47 and K = .87), and the lower values resulted from slightly poorer agreement between judge b and judges a, c, and d, than the agreement amongst the other judges.

Table 4 - Kappa scores for the four validation study judges, a, b, c, and d. Scores are based only on stimuli not used in the original study.

	a	b	c	d
a	-	.53	.80	.87
b	.53	-	.67	.47
c	.80	.67	-	.73
d	.87	.47	.73	-

Table 5 - Kappa scores for the original filters, F, and the four validation study judges, a, b, c, and d.

	F	a	b	c	d
F	-	.10	.08	-.03	-.07
a	.10	-	.30	.77	.77
b	.08	.30	-	.27	.33
c	-.03	.77	.27	-	.80
d	-.07	.77	.33	.80	-

Table 5 shows kappa values between all four validation study judges, and the original filters, F, and is based only on the 60 stimuli used in the original data reduction procedure. Looking only at comparisons between judges a, b, c, and d, kappa ranged between K=.27 and K=.8, indicating that there was generally good to moderate agreement between those judges. The lower values of kappa again resulted from relatively poor agreement between judge b and judges a, c, and d.

However, Table 5 also shows that there was very poor agreement between each of the four validation study judges and F (ranging between K = -.07 and K = .10). A further analysis revealed that the cases of disagreement resulted largely from the four judges counting many cases of extreme overlap as responses, where the original filters would not have. The percentage of sample cases which were judged as non-responses and remained after the filters had been applied (in other words, the 'noise' in the cases which remained in the analysis) were 16.7% by F, and 9.6% by the judges' decisions. The percentage of sample cases which were judged as responses but which had been eliminated by the cut-off points (in other words, valid cases which nonetheless were left out of the analysis) were 0% by F, and 33.5% by the judges' decisions. Tables 6-9 show the agreement between each judge and F. These show how the validation study judges tended to disagree most with F concerning those cases excluded by F. Judges a-d decided that many such cases were responses, when F decided they were not.

Table 6 - Agreement scores between judge a and F.

judge a	F			
	included by filters		excluded by filters	
	response	non-response	response	non-response
response	10	7	0	20
non-response	0	3	0	20

Table 7 - Agreement scores between judge b and F.

judge b	F			
	included by filters		excluded by filters	
	response	non-response	response	non-response
response	3	5	0	15
non-response	7	5	0	24

Table 8 - Agreement scores between judge c and F.

judge c	F			
	included by filters		excluded by filters	
	response	non-response	response	non-response
response	8	8	0	21
non-response	2	2	0	19

Table 9 - Agreement scores between judge d and F.

judge d	F			
	included by filters		excluded by filters	
	response	non-response	response	non-response
response	9	7	0	24
non-response	1	3	0	16

4.3.4.3 Conclusions

In conclusion, the cut-off points set by the original judgements eliminated cases that the other judges classed as responses. But in so doing the original cut-off points filtered out more 'noise' from the final analysis, because the proportion of non-response cases let through by the filter would have been lower (assuming that the proportion of non-response cases to response cases remains approximately equal as the cut-off points become less stringent - an assumption suggested by the histograms in the preceding sections). Accepting the original filters therefore provides stricter filters, and a more conservative data set than would have been the case if the validation study judges' assessments had been used. Since this conservative approach is generally preferable, the decision was taken to accept the original filters.

4.4 Determining Response with Positive Intervals

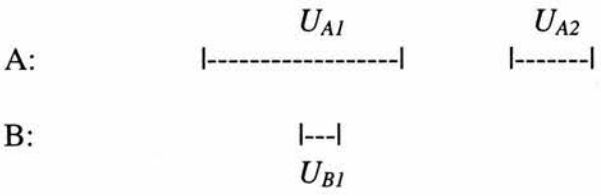
Many inter-speaker intervals in the Map Task Corpus may arise not because of linguistic or cognitive factors, but because the follower has to draw part of the route on the map. That is, speaker B may only respond to A's utterance verbally once he or she has responded physically. The length of the interval is therefore determined to a large extent by how long it takes for the route to be drawn. One might argue that sometimes B's response comes before the drawing is finished, and that therefore could represent some genuine linguistic or cognitive response time. But in most cases

it appears that an utterance is produced only after drawing is finished. Other intervals were largely filled with laughter or coughing or other extraneous noises. Again, it could be argued that the length of such intervals results from how long one or both of the speakers laughs or coughs for.

On the other hand, there is no ready means to assess whether an exchange involves some delay other than by examining all exchanges with a positive interval. A preliminary analysis of 25 exchanges with inter-speaker intervals greater than 1000ms revealed no relationships between interval duration and other activities during the interval. The problem here is that there is no simple way of categorising these cases in a rule-based or time-based way because some relatively short positive intervals could involve some amount of drawing.

A further analysis of 6 intervals which were greater than 5000ms revealed that while two involved some drawing, two appeared to be the result of uncertainty or confusion on the part of one or both of the participants. The other two were caused by a problem demonstrated in 26) below.

26)



Here, using the start points of each utterance, the order of these three utterances is U_{A1}, U_{B1}, then U_{A2}. This results in overlap between U_{A1} and U_{B1}, and a positive interval between U_{B1} and U_{A2}. But if U_{B1} were heavily embedded within U_{A1}, it seems doubtful whether U_{A2} would really be a response to it, and whether U_{A2} might not be better treated as a continuation of U_{A1}.

There does not appear to be any satisfactory solution to this problem, because it is quite conceivable that U_{A2} could be a genuine response to U_{B1}, so making the interval valid. For this to be the case, U_{B1} may have to be uttered relatively near the end of U_{A1}. It would have been possible to use this consideration for some individual

cases, but such results could not have been extrapolated to all cases of positive interval.

Further to this, an analysis of positive inter-speaker intervals greater than 0ms and less than 100ms was carried out. There were 1425 intervals within this range. Every 20th exchange was taken from these, giving a sample of 72 cases. The same criteria were used as for the assessment of the negative intervals.

Results from this analysis showed that of the 72 cases, 22 (30.56%) were non-responses and 50 (69.44%) were responses. It seems likely therefore that there would continue to be instances of non-response until a somewhat higher positive interval. This is particularly so when one considers that moves starting new games, and moves following backchannel signals (that is, non-response type moves) may conceivably follow on from a previous utterance by several seconds.

Because of these difficulties with testing the validity of exchanges with positive intervals, no definite pattern could therefore be established. Unfortunately the presence of utterances which were most likely not responses has had to be tolerated for the purposes of this research.

4.5 Sub-move Units

An added factor was that even where U_B could be treated as a response to some specific move within U_A, it may in fact be a response to an earlier *part* of that move.

Take example 27) below. In move 116, the follower uses a *query-yn* to find out how far to draw a line. The giver uses a *reply-n* in move 117, and there seems little doubt at this stage that it is a response to move 116. For one thing, a *query-yn/ reply-n* move combination seems like a valid pattern for an exchange. Also, no 'cont' feature is present on move 117, and nor does it start a new game. But what exactly is move 117 responding to? Example 27) above shows that the giver's move starts 0.82 seconds after the word "*mine*", and 0.03 seconds *before* '*or*'. From this it seems fairly clear that the giver can only be responding to that part of the follower's utterance preceding the pause.

27)

follower move 116 *query-yn*

START=195.8294 **the**

START=195.9065 **diamond_mine**

START=196.4019 PAUSE[0.8224]

START=197.2453 **or**

START=197.3201 **the**

START=197.3832 **banana**

START=197.7570 **tree**

giver move 117 *reply-n*

START=197.2211 **no**

START=197.5703 **no**

START=197.7717 **not**

START=197.9073 **as**

START=198.0717 **far**

START=198.2813 **a--**

START=198.3841 <unknown>

START=198.4621 **not**

START=198.6306 **f--**

START=198.7341 <unknown>

But note that move 116 is labelled as being just one move. From a functional point of view this is valid, because both the part preceding and the part following the pause in 116 have the same basic function - that of asking a y/n-question. But from a dialogue-based point of view this analysis is clearly inadequate.

Cases such as this were not uncommon - in a provisional analysis of 70 overlapped exchanges, 37 were found to contain second moves which were responses to some earlier move in an utterance, or to an earlier part of a move. It seemed that even the move unit was too large a unit, and that the responses in these cases seemed to be to some smaller unit - perhaps a TCU - within the move. As I suggested earlier, the basic unit of turn-taking may be the TCU, and the move may be too simplistic to act as a basic unit itself. Because prior sub-classification had not been made it was not possible to calculate response to any finer degree than to the nearest move. So in the example above, overlap would be calculated between the start of the giver's move

and the end of the follower's whole move, rather than the more accurate measurement from the end of the word '*mine*' to the start of the giver's move (which here would result in a positive interval).

4.6 Summary

This chapter has described the basic units of analysis used in this research. Turn units have been rejected in favour of move units. This was because it was felt that move units have advantages over turn units both practically and theoretically. The definition of an utterance, as used for this research, is based on splitting speech into moves, and taking as the boundaries of an utterance those points where a speaker switch occurs.

However, there are two basic problems with an analysis of conversation: a) the problem of deciding whether an utterance by a next-speaker (N) is a response to an utterance by the current-speaker (C), and b) if it is a response, there is the problem of deciding which *part* of the utterance it is responding to.

These problems are important because not all conversation is made up of an orderly pattern of turn-taking, where each utterance is a response to the immediately preceding one by the other speaker. An utterance may start a new topic of conversation, or it may take place in the back-channel and therefore not compete for the conversational floor. Utterances may occur simultaneously, or almost so, with one another, or they may overlap by considerable amounts. In an analysis of the temporal coordination of utterances a description is needed of which utterances are being coordinated. And if an utterance B overlaps an utterance A by several seconds, one is forced to question whether the start of B is being coordinated with the end of utterance A. According to Clark's theory the overlap should signal something about the mental states of the speaker - for example, an intention to take the floor because the speaker has heard enough for current purposes to make a contribution, or a miscalculation of a TRP. Or it may reflect certain social factors (common ground) between the two speakers. Unfortunately, in the case of overlap it is not always obvious whether the overlap is an incorrect anticipation of an imminent TRP, or is a

response to a part of an utterance which came earlier (and is therefore not an overlap at all).

The method used here to try to solve these problems was to analyse individual exchanges and test them for response determination and response location. Several criteria were used, including use of the move, feature and game coding in the Map Task data. Also, moves following backchannelled moves were not counted as responses. Because the corpus contained over 18,000 exchanges, a full analysis of each exchange was not feasible. A sample was therefore used, and a cut-off point established for three types of interval in the exchanges. These intervals could be calculated for all exchanges, and those falling below the cut-off points were omitted from the analysis.

The initial data set contained 18,248 exchanges. Table 10 shows how this original set was reduced by 2805 exchanges. A further 1546 exchanges were eliminated purely through application of the automatic filters, and 1169 exchanges were eliminated through misclassifications or errors in the original coding.

Table 10 - Reduction of the original data set (n = 17079) with the application of each filter, where filter 1 is the excessive overlap criterion, filter 2 is the near-simultaneous criterion, and filter 3 is the uncoded continuation criterion. The term 'Automatic filters' refers to the application of the 'cont' feature and backchannelling constraints (as per 4.4.1). 'Subjective filters' refers to filters which relied on examination of semantic content, move and game coding, and the results of analysis of samples of data (as per sections 4.4.2 and 4.4.3)

Filter	Cases eliminated by automatic and subjective filters	Cases eliminated by subjective filters alone	Total eliminated
Filter 1 only	29	395	424
Filter 2 only	624	567	1191
Filter 3 only	129	238	367
Filters 1 and 2	53	277	330
Filters 1 and 3	1	24	25
Filters 2 and 3	235	149	384
Filters 1, 2, 3	23	61	84
Total of all filters	1094	1711	2805

5. Analyses of the Rhythmic Coordination Hypothesis

5.1 Introduction

Despite some reasonable theoretical arguments for rhythmic coordination in turn-taking (Couper-Kuhlen, 1993), there still remains little reliable empirical evidence. Couper-Kuhlen carried out an analysis of conversational data which relied on two trained listeners to assess whether perceptually isochronous sequences existed within and across utterances. This assessment was made through tapping along to the conversation - a process which has been claimed to aid the perception of rhythm (cf. Allen, 1972; Donovan & Darwin, 1979; Darwin & Donovan, 1980). But there are two points here. First, are the judgements of two trained listeners representative? Second, and perhaps more important, is the fact that tapping was necessary to facilitate the perception of isochrony, and that judgements were made after considerable deliberation. In real conversation, participants must be able to make extremely fast assessments of the isochronous or non-isochronous nature of speech. Presumably, the perception of isochrony would be subconscious. Therefore, a test of isochrony at a conscious level is problematic. What are required are perceptual experiments which indirectly test the assumption that isochronous turns which are rhythmically coordinated are the unmarked case - or appear in some sense more 'natural' than non-isochronous utterances.

This chapter presents two data analyses, where mean inter-stress intervals in one participant's speech were compared with inter-stress intervals across speaker boundaries (the *between-interval*). In the pilot analysis, a small subset of the corpus was labelled for prominent syllables by hand. In the second analysis, the whole corpus was labelled using an automatic stress assignment model. These analyses

worked from the hypothesis that if isochronous sequences are as common as they have been reported to be, and if N's first stressed syllable should be perceived to fall on the first, second or possibly third beat after C's last stressed syllable, then the ratio of between interval to mean inter-stress interval (the *rhythmic ratio*) should be *approximately* equal to an integer. Of course, it has not been suggested that perceptual isochrony is the same as acoustic isochrony. But Couper-Kuhlen (1993) does give an account of the acoustic 'window' within which a perceptually isochronous sequence must fit. It was therefore expected that for the rhythmic hypothesis to hold, a large number of exchanges should have a rhythmic ratio which was an integer, plus or minus a degree of freedom to account for the discrepancy between perceptual and acoustic isochrony.

In the event that the experiments and analyses did not provide any evidence of cross-speaker isochrony in real conversation, it will be concluded that it is not used as a major strategy in the coordination of turn-taking.

5.2 Perception of Differences in Inter-stress Intervals between Speakers

5.2.1 Introduction

The rhythmic coordination hypothesis assumes that it is the *perception* of isochrony which acts to coordinate conversation. It also predicts that speaker switches which contain isochronous sequences are the unmarked case, and should be perceived to be more 'natural' exchanges. It was therefore felt necessary to test whether subjects are able to listen to recordings of exchanges, and to alter the between interval until the interval sounds the most 'natural'. If there is any validity to the rhythmic hypothesis, then subjects should tend to alter the between interval to maintain the perceived isochrony.

If listeners cannot reliably differentiate 'normal-sounding' between-intervals from between-intervals which sounded 'too short' or 'too long', then isochrony is unlikely to be an issue. As the results from this pilot study indicate, subjects were able to carry out such a differentiation task in most cases, and there are grounds for

continuing with the experiment to test the rhythmic turn-taking hypothesis as outlined above.

5.2.2 Method

In a pilot study ten subjects were presented with five exchanges taken from the Map Task Corpus. I selected these utterances on the basis that the transitions from one speaker's utterance to the next speaker's utterance did not involve hesitations, false starts, or other disfluencies, and were not overlapped. I used the original inter-speaker interval for each exchange, and then three other versions where the inter-speaker interval had been doctored: one with a between-interval of half the duration of the original (the short version), another with one and a half times the length of the original (the long version), and another with double the duration of the original-interval (the very long version). This gave four versions for each exchange - a total of twenty stimuli. I presented the subjects with all twenty stimulli in random order, and then again in a different random order. The subjects had to decide for each stimulus whether the between-interval appeared to be of a natural length, longer than natural, or shorter than natural.

5.2.3 Results

The relationship between the choices made by the subjects in the first random sample of exchanges and choices made in the second presentation of exchanges were not significantly different ($\chi^2 = 65.61$, df = 4, $p << 0.001$).

For all five exchanges there were significant differences between the inter-speaker interval categories of normal, shorter than normal, and longer than normal ($\chi^2 = 90.61$, df = 6, $p << 0.001$). These categories were significantly different amongst both Map Task familiar subjects and Map Task unfamiliar subjects (for map-task-familiar subjects n = 160, $\chi^2 = 33.23$, df = 6, $p << 0.001$; for map-task-unfamiliar subjects n = 240, $\chi^2 = 64.16$, df = 6, $p << 0.001$). Subjects, however, tended to choose the "normal" category more than the other categories (in 52.5% of cases, as opposed to 25.5% for "long" choices, and 22% for "short" choices.

Figure 1 below shows the relationship between the duration of the stimulus between-interval and the estimates of "short" and "long" for each of the four stimulus categories (each stimulus category is represented by a different point on the plot. Note that the lines in the plot are used purely to clarify the relationship between different exchanges, which are numbered from 1-5). One would expect all the plots to follow similar patterns if there were a simple shift in the perception of "short" and "long" categories. But this is not evident.

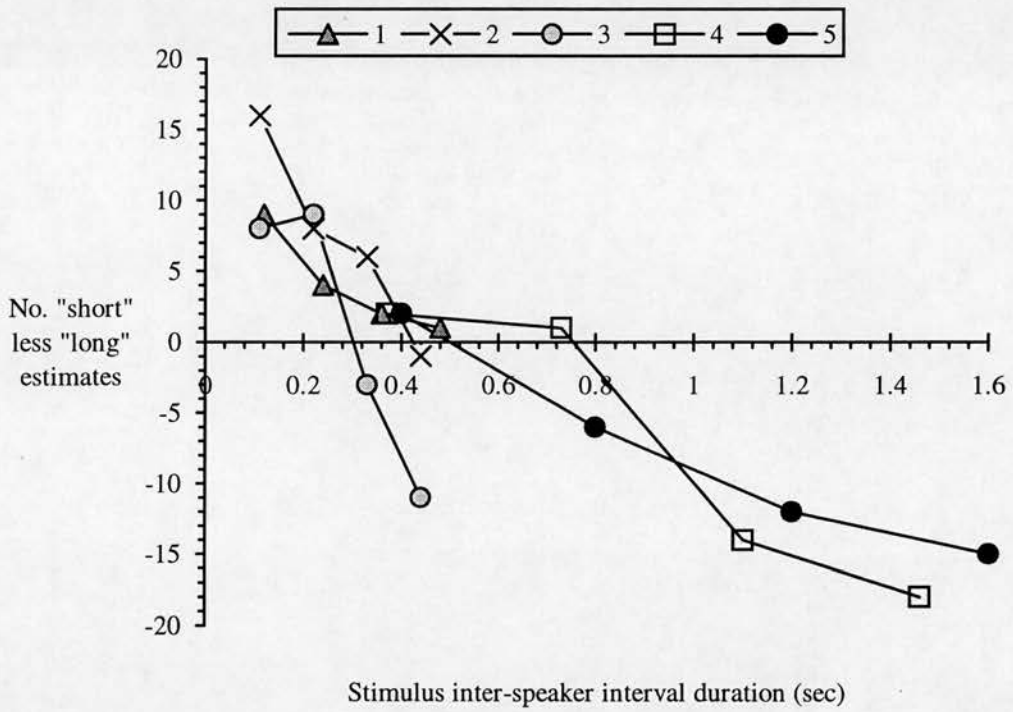


Figure1. The number of "short" judgements less the number of "long" judgements as a function of stimulus between-interval, for each of the five exchanges.

5.2.4 Discussion

The results suggest that subjects were able to distinguish between lengthened and shortened stimulus between-intervals. They tended to prefer to choose "normal" for many stimuli, even when presented with the doctored versions. That is, subjects very often considered the stimulus interval to be of a 'natural-sounding' duration irrespective of its actual duration. There must be considerable tolerance for what appears to count as a normal length of an between interval in dialogue. This might,

however, reflect subjects' general tendency towards caution in making their judgements.

Related to this is the observation that different types of exchange lead to different responses, despite identical or similar natural between intervals. That is, the context of utterance appears to play a role in the between intervals that are expected, or would be considered unmarked.

The original between-interval was also a factor, with subjects scoring more accurately on their "long" decisions when the original-interval was long than when it was short. When the original between interval was relatively long, subjects tended to score "normal" more often for unaltered and shortened stimulus intervals than in the exchanges where the original-interval was relatively short (see Appendix D for further details). When the original between-interval is longer, subjects are better able to discriminate between different intervals.

5.3 Preferred Between-Intervals and Isochrony

5.3.1 Introduction

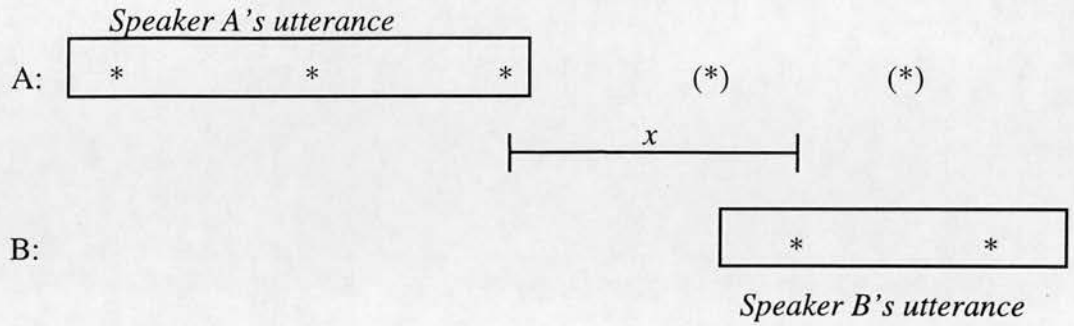
The pilot study provided provisional evidence that subjects could distinguish among between-intervals. I carried out two further experiments to see if subjects could agree on preferred between-interval durations.

The rhythmic coordination hypothesis (see chapter 4) claims that participants in a conversation perceive speech, both within and across utterances, as isochronous sequences of syllables - that is, they perceive syllables to be approximately equally separated in time. The hypothesis predicts that the unmarked case for speaker switch is a perceptually isochronous sequence of syllables across utterances.

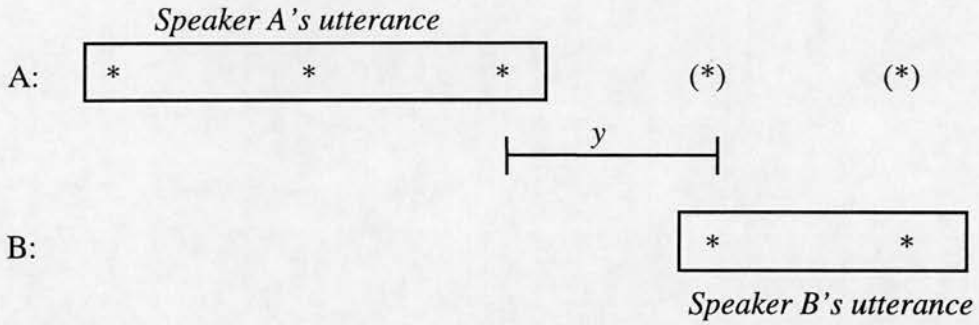
If subjects were presented with recordings of exchanges from conversations where the between-intervals were of varying durations, and were asked to make judgements about which exchanges appeared more 'natural' (or unmarked), the rhythmic coordination hypothesis predicts that they should choose those exchanges where an isochronous sequence was carried over a turn transition. Further, if subjects

were able to manipulate the durations of the between-intervals, they should alter them to produce isochronous sequences across turn transitions.

1a)



1b)



In example 1a), the first stressed syllable of speaker B's utterance, U_B (marked by an asterisk) does not coincide with the beat set up by the isochronous sequence in speaker A's utterance, U_A (the 'imaginary' beat is expressed by a series of asterisks in parentheses). Example 1a) therefore expresses a marked exchange. Subjects should judge cases like example 1a) to be less natural than 1b), where the first stressed syllable of U_B does coincide with the imaginary isochronous sequence. Given the option to alter the duration of the interval from x to y ms, a subject should change 1a) to 1b).

The duration of the interval in an isochronous sequence (y ms in example 1b) should be equal to, or some multiple of, the mean inter-stress interval in U_A. The target interval may be some multiple of A's mean inter-stress interval because

Speaker B's first stressed syllable need not coincide with the first imaginary beat after the end of speaker U_A. It may coincide with the second or third imaginary beat.[17] Note also that the rhythmic coordination hypothesis works under the assumption that isochrony is a perceptual notion. Therefore, if subjects were to alter between intervals to produce isochronous sequences, it should not be expected that the intervals between stressed syllables would be exactly equal in duration. However, Couper-Kuhlen (1993) does provide some evidence for the relationship between perceived and acoustically measured isochrony. She claims from empirical analysis (see chapter 2) that the perception of an isochronous sequence breaks down when an inter-stress interval varies by ±20-30% from a previous inter-stress interval. For the purposes of this experiment, perceptual isochrony was therefore considered likely across a turn transition if the inter-stress interval across a turn transition differed by no more than 30% of the duration of the previous inter-stress interval.[18]

Results from both experiments reported below failed to support the rhythmic coordination hypothesis, however. Instead, there appear to be far more contextual factors involved than can be accounted for by a simple rhythmic principle, and isochrony is at best one of several possible cues which help coordinate turn-taking.

---

[17]It is not certain what upper limit there could be to the number of imaginary beats that could pass before the next-speaker starts his or her utterance, and yet for a perceptually isochronous sequence still to hold. Couper-Kuhlen (1996) has suggested an upper limit of three or possibly four, although it is not clear what basis she has to make this claim.

[18] Although Couper-Kuhlen used a figure of 20-30% of the pervious inter-stress interval, it is also possible to basis the margin of freedom on the mean of several previous inter-stress intervals.

5.3.2 Experiment I

5.3.2.1 Method

I selected twenty-five exchanges, each consisting of an utterance by one speaker lasting a few seconds, an interval, and an utterance by another speaker, from the Map Task Corpus. The exchanges chosen did not contain any major disfluencies, hesitations, or false starts. The between-intervals in the original recordings (hereafter called *original-intervals*) were relatively evenly distributed between 0 and 1000ms.

The original-intervals for each exchange were then altered, being replaced with intervals between 0 and 1000ms, generated randomly. The artificially generated intervals consisted of low-volume 'noise', and the amplitude of 20ms of the speech either side of this noise was acoustically tapered to prevent a noticeable click as the signal switched from speech to noise and back to speech.

The altered exchanges were presented to two groups of subjects - referred to as Group A and Group B respectively. Each group consisted of fifteen subjects. The start-intervals (the between-intervals that the subjects heard initially) of the presented exchanges were significantly different for each group. The exchanges were presented to each subject in a randomly generated order (the ordering was the same for each subject), over headphones. A transcript of the exchanges was provided to help comprehension in case subjects had difficulty understanding any exchange. Subjects had to press a key to start the playback of each exchange, and could listen to it as often as they wanted. A second key had to be pressed when they were ready to proceed to the next exchange.

The subjects were instructed to listen to each presented exchange and to modify the between-interval from the duration they first heard until they felt that the interval was of a duration that it would be in a natural discourse environment (this final interval is hereafter referred to as the *finish-interval*).

Subjects were able to make the interval duration greater or smaller by increments of 50ms or 150ms by pressing appropriate marked keys on a keyboard. Each time a subject pressed a key, the entire exchange was played again, with the altered between interval. Subjects could make as many adjustments as they felt necessary to increase or decrease interval duration. They were also able to listen to

the exchange again without altering the interval duration from its previous presentation.

5.3.2.2 Results

I carried out a multiple regression analysis, indicating that a significant proportion of the variance in finish-intervals could be accounted for by an equation including start-intervals and original-intervals (R^2 = 0.3766, F (2, 747) = 225.673, $p < 0.0001$). While longer start-intervals made for longer finish-intervals ($\beta = 0.61$, $p < 0.001$), original-intervals had no significant effect ($\beta < 0.01$, $p = 0.98$). That is, it appeared that subjects were altering the between intervals in relation to the interval present on the first presentation of an exchange, rather than in relation to the interval present in the original dialogue. Figure 2 below shows how as the start-intervals increase, the mean finish-intervals also increase. The finish-interval is to a large extent dependent on the start-interval.

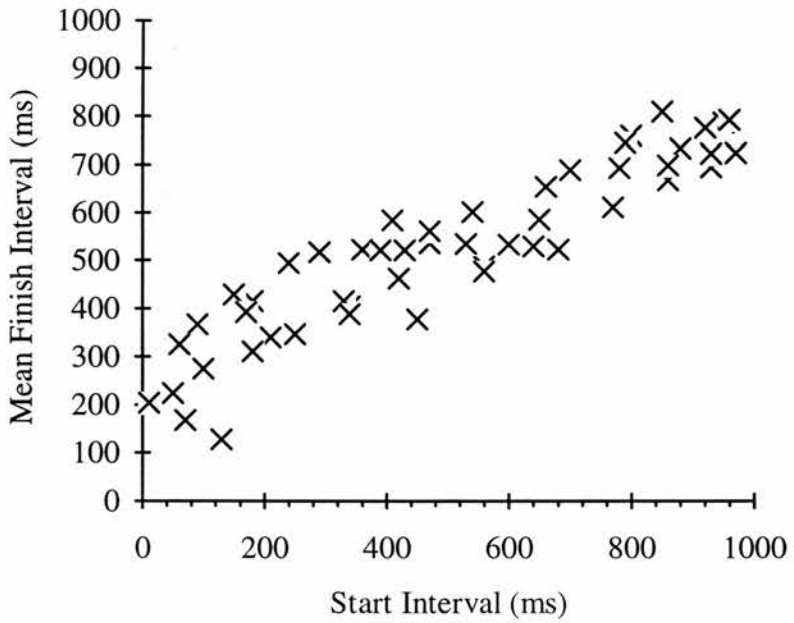


Figure 2 - Experiment I. Plot of start-interval against Mean finish-interval

I made a comparison between the choices made by the two groups of subjects. Each group was presented with exchanges which had a significantly different set of start-

intervals (r = 0.02, p = 0.924). One would have expected this difference to have been eliminated if subjects were to rely on a common rhythmic mechanism for determining between interval durations. A simple regression analysis of the relationship between the finish-intervals of Group A and the finish-intervals of Group B showed no significant correlation (r = 0.035, p = 0.505), as shown in Figue 3 below. This indicated further that start-intervals had a significant bearing on the results. It also indicated that subjects were not using any form of perceptual isochrony in their decisions. Decisions based on perceptual isochrony should operate independently of the start-interval, and should allow subjects from the two groups to produce finish-intervals that correlated.

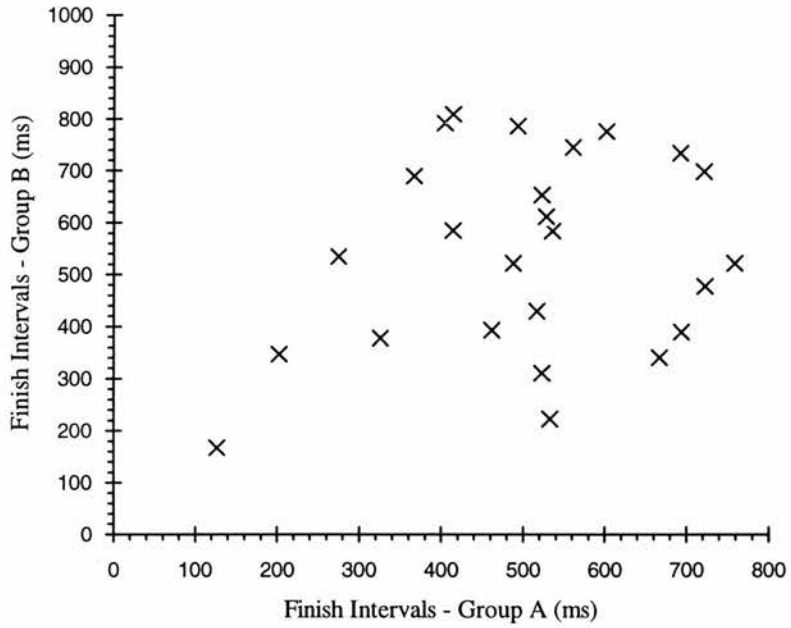


Figure 3 - Experiment I. Plot of finish-intervals of Group A against finish-intervals of Group B

Finally, finish-intervals tended to be less variable than the start-intervals (start-interval mean = 499.2ms, sd = 293.75ms; finish-interval mean = 521.06ms, sd = 248.93ms). In fact, the more the start-interval is near 0ms or near 1000ms, the bigger the change made. Subjects therefore tended to choose finish-intervals smaller in

range than the start-intervals were, and hence converged on some 'average' interval (see Figure 4).

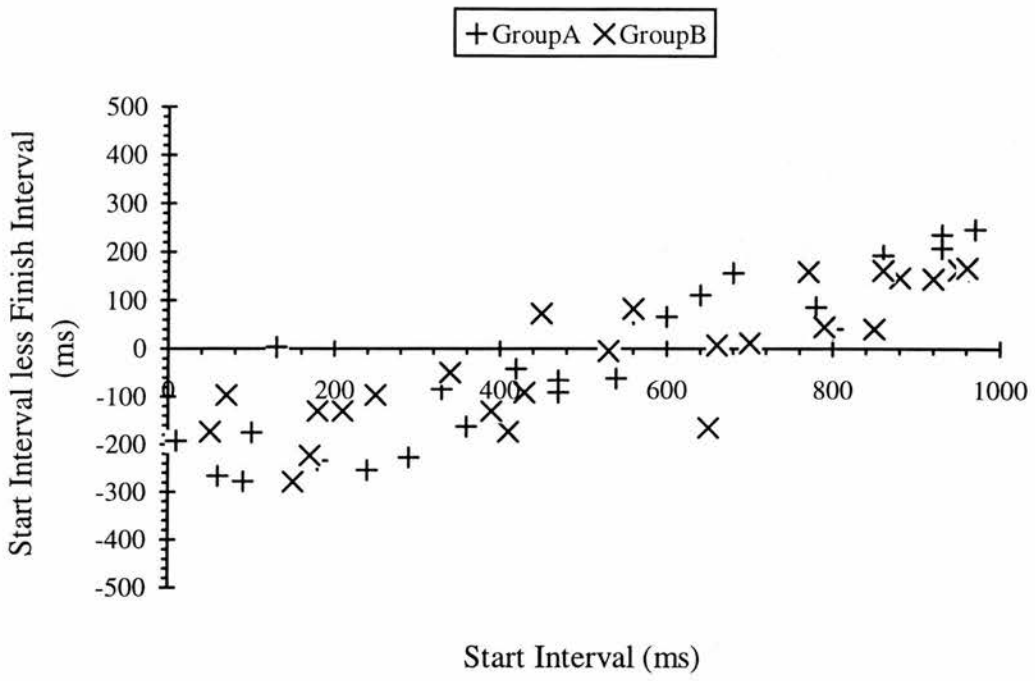


Figure 4 - Plot of start-interval against start-interval less finish-interval

5.3.3 Experiment II

5.3.3.1 Method

The aim of Experiment II was to see if dialogue context had any effect on the task in Experiment I. Instead of hearing only one short utterance or part of an utterance, the between interval, and then an utterance by the other participant, subjects were presented initially with approximately ten seconds of speech either side of the turn transition. A transcript of each exchange was provided. The turns immediately surrounding the target transition point were highlighted, so that subjects knew which turn transition in the dialogue to pay attention to. Subjects heard a reduced version of the exchange on subsequent presentations, as in experiment I. Note that the start-interval in both the first presentation of the dialogue and the first presentation of the target exchange were identical. At no stage was the original-interval presented.

The methodology for experiment II was otherwise identical to experiment I. The same twenty-five exchanges were used, and were presented to the subjects in the same random order as in experiment I. Again, there were two groups of fifteen subjects, who had control of when and how many times they could listen to each exchange, and how many times the between interval could be altered. The increments of each alteration were either 50ms or 150ms, depending on which key the subject pressed.

5.3.3.2 Results

I carried out a multiple regression analysis, which again indicated that a significant proportion of the variance in finish-intervals could be accounted for by an equation including start-intervals and original-intervals ($R^2 = 0.4519$, $F\ (2, 747) = 307.919$, $p < 0.0001$). Longer start-intervals produced longer finish-intervals ($\beta = 0.65$, $p < 0.001$). Original-intervals now had a very small, yet significant, influence on finish-intervals ($\beta = 0.07$, $p = 0.01$).

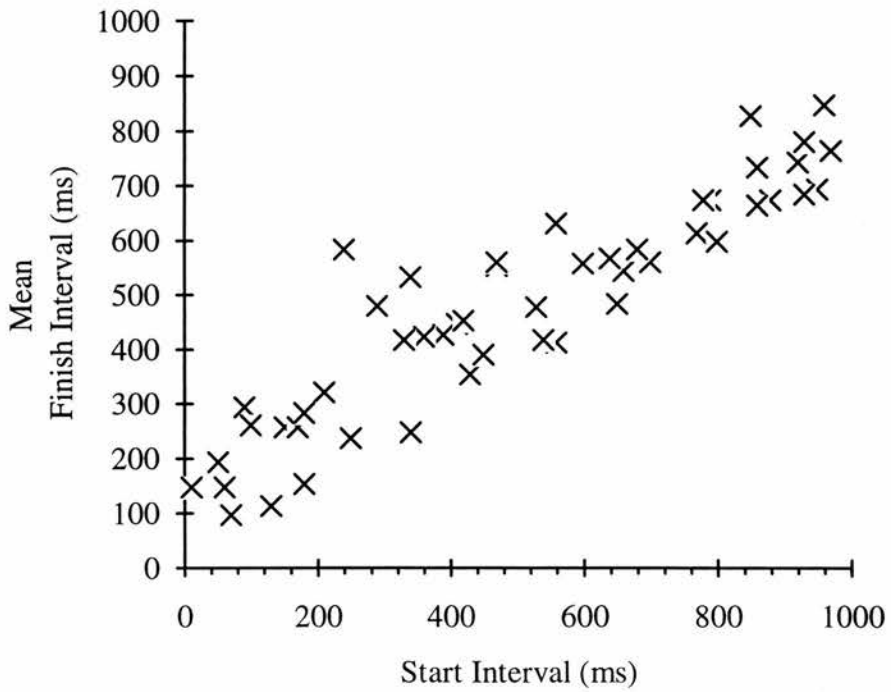


Figure 5 - Experiment II. Plot of start-interval against mean finish-interval

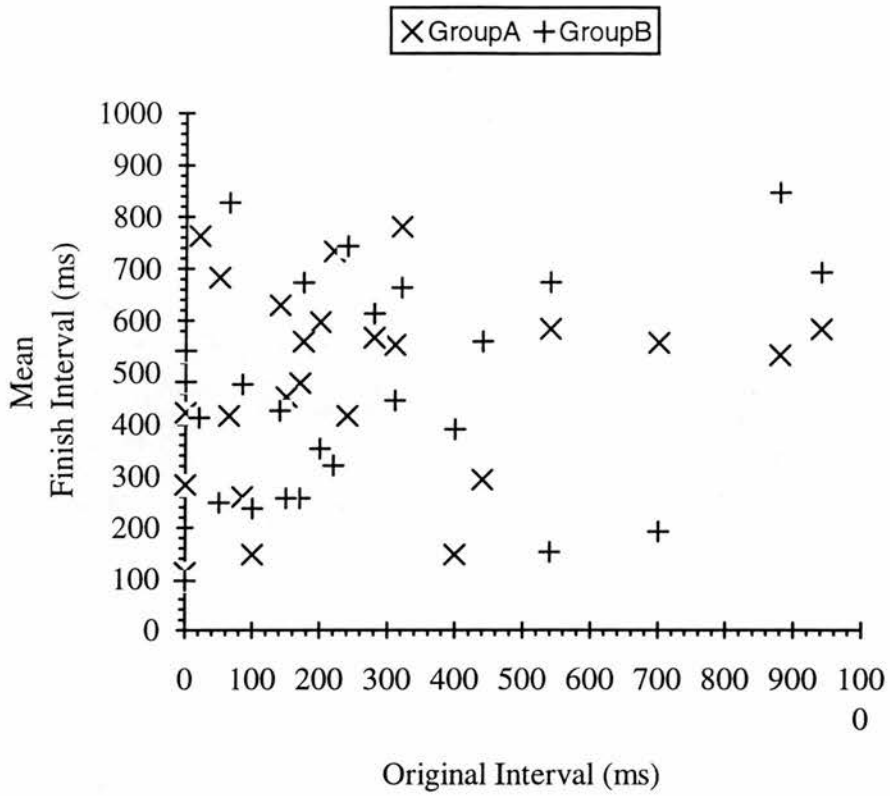


Figure 6 - Experiment II. Plot of original-interval against mean finish-interval

Given greater context, subjects were influenced to a small extent by the original-intervals (which they never encountered). This suggests that additional context may yield more or better cues to some idealised interval.

I made acomparison between the choices made by the two groups of subjects. A simple regression analysis of the relationship between the finish-intervals of group A and the finish-intervals of group B was just short of significance ($r = 0.1$, $p = 0.053$). Therefore, while this result indicates as in experiment I that start-intervals must have had a significant bearing on the choice of finish-intervals, it also supports the findings from the multiple regression analysis that original-intervals had a greater influence than in experiment I.

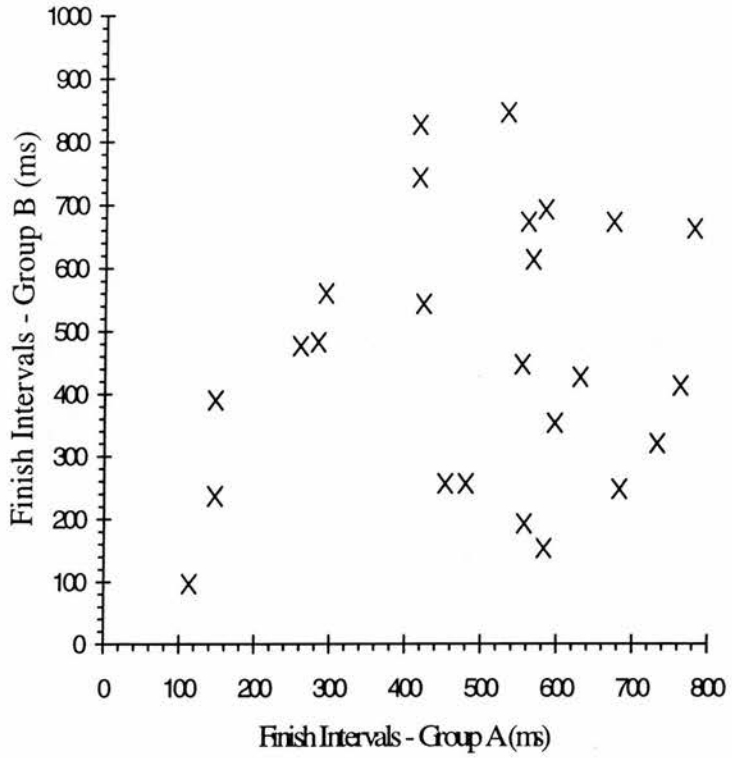


Figure 7 - Experiment II. Plot of finish-intervals of Group A against finish-intervals of Group B

As in experiment I, finish-intervals were less variable than the start-intervals (start-interval mean = 499.2ms, sd = 293.75ms; finish-interval mean = 476.38ms, sd=265.34ms). Again, more extreme start-intervals were subject to more alteration ($r = 0.498$, $p < 0.001$). So, as in experiment I, subjects chose finish-intervals which were regressed towards a 'central' value.

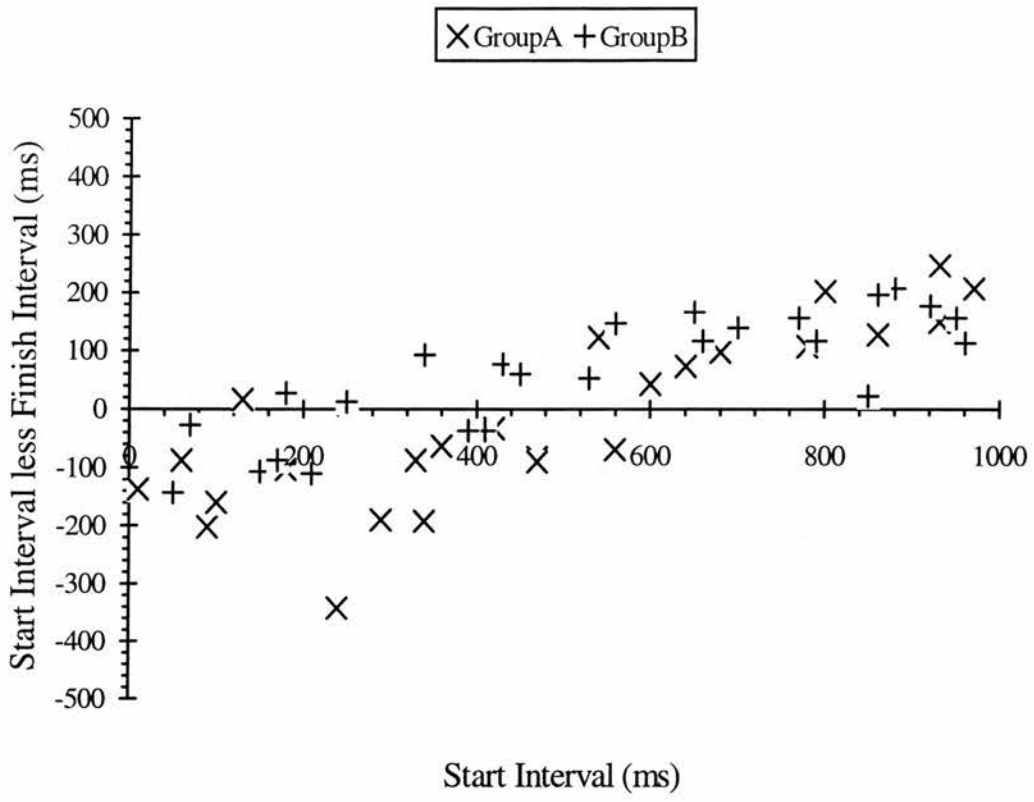


Figure 8 - Experiment II. Plot of start-interval against start-interval less finish-interval.

5.3.4 Results from a Comparison of Experiments I and II

The mean finish-intervals in experiment I correlated strongly with mean finish-intervals in experiment II ($n = 50$, $r = 0.916$, $p < 0.001$). This is shown in Figure 9 below. Although subjects in different groups *within* each experiment did not agree on their finish-intervals, there was agreement in the finish-intervals amongst subjects in corresponding groups in different *experiments*.

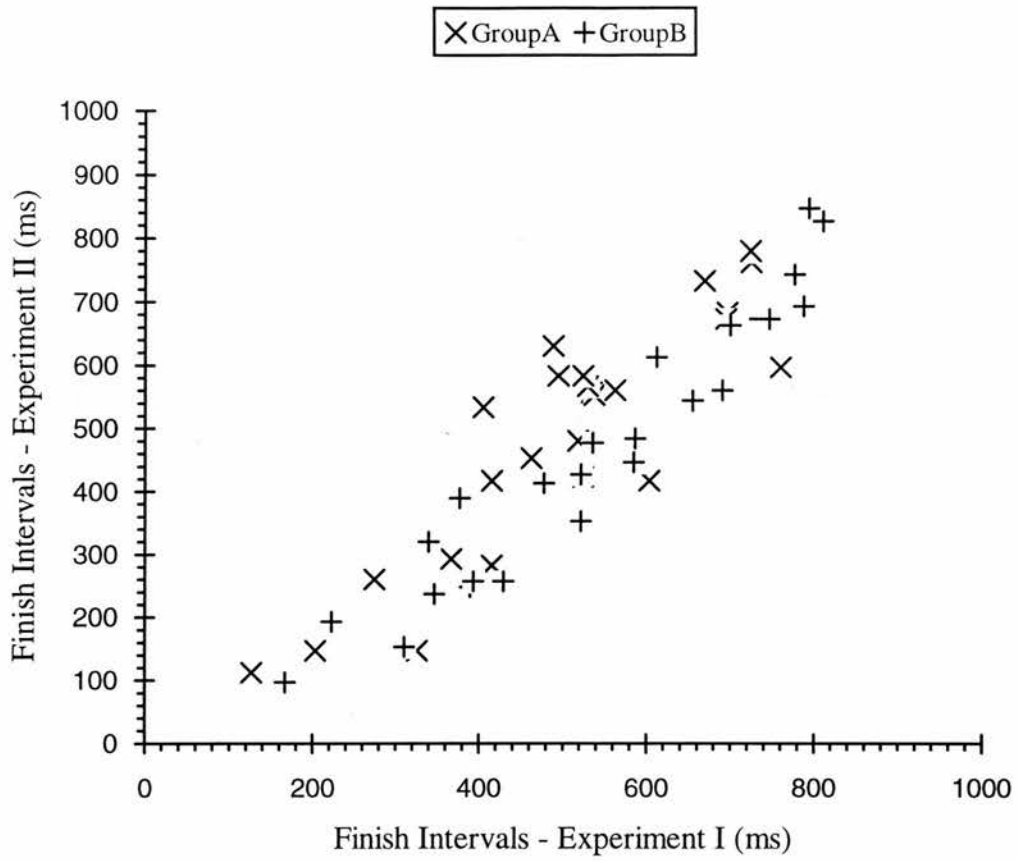


Figure 9 - Plot of Mean finish-intervals in Experiment I against Mean finish-intervals in Experiment II

5.3.5 General Conclusions for Experiments I and II

Experiments I and II led to the conclusion that subjects were primarily influenced by factors other than perceptual isochrony to judge the 'ideal' interval in exchanges. In both experiments different groups of subjects were presented with similar stimuli - the only variable being the duration of the between interval that they were initially presented with. If a theory of perceptual isochrony and its unmarked discourse function were to hold, the different groups of subjects should alter the intervals to similar final durations. However, they did not. Instead, there appeared to be a marked tendency for subjects to choose between intervals which did not differ vastly from the between intervals that were initially presented to them. This seemed to indicate that subjects were often not sufficiently sensitive to differences in between-intervals that

they felt they needed to alter the interval greatly. The results simply show a regression towards the mean.

Because of the between-group disparities in both experiments, it was not felt necessary to continue with an analysis of possible isochrony in the exchanges. This would only have been necessary if some correlation had been found in the results from different groups of subjects.

Experiment II did reveal a very small significant relationship between original-intervals and finish-intervals, indicating that access to more contextual information can affect the choice of finish-interval. This finding is not surprising if one assumes that context plays an important role in all aspects of language. The problem here is to decide whether additional context gives subjects a better sense of the rhythmic structure of a dialogue. In theory, the only rhythmic structure that ought to be necessary are the few beats occurring before the turn transition. These were present even in the low-context situation in experiment I, where original-intervals had no significant effect on finish-intervals.

One possible objection to these findings is that subjects, rather than altering the intervals freely until they were completely satisfied that they had reached a 'natural' interval, felt that they were under some time pressure and chose a value which was not vastly different from the start-interval. Anecdotal evidence from the subjects would suggest however that this was not the case.

Also, an explanation for the lack of correlation between original-interval values and finish-interval values may have been caused by the non-spontaneous nature of the task. That is, that subjects were aware of an upcoming turn transition point through repeated exposure to an exchange, whereas in natural dialogue hearers would possibly be able to detect a turn transition point before it occurred through syntactic, pragmatic, intonational or other cues (see chapter 2). Of course, if a rhythmic process were being used in natural dialogue, the results ought not be affected by repeated exposure, since in both natural and artificial situations the same rhythm would be timing the start of the second turn. But overriding these considerations is the observation that in both experiments the original-interval was at best significantly less of a factor than the start-interval.

While there were significant correlations between start-intervals and finish-intervals, there was evidence that longer start-intervals produced finish-intervals which were slightly shorter, and vice versa. Subjects seemed to be sensitive to different intervals, in that they were able to detect whether a given start-interval was particularly long or short. They then attempted to adjust the interval on this basis. These experiments therefore support the findings of the pilot experiment - namely, that there is a relatively broad tolerance of what constitutes a 'natural' interval, though subjects were able to discriminate long from short intervals.

5.4 Analysis of Map-Task Data: manually-labelled data

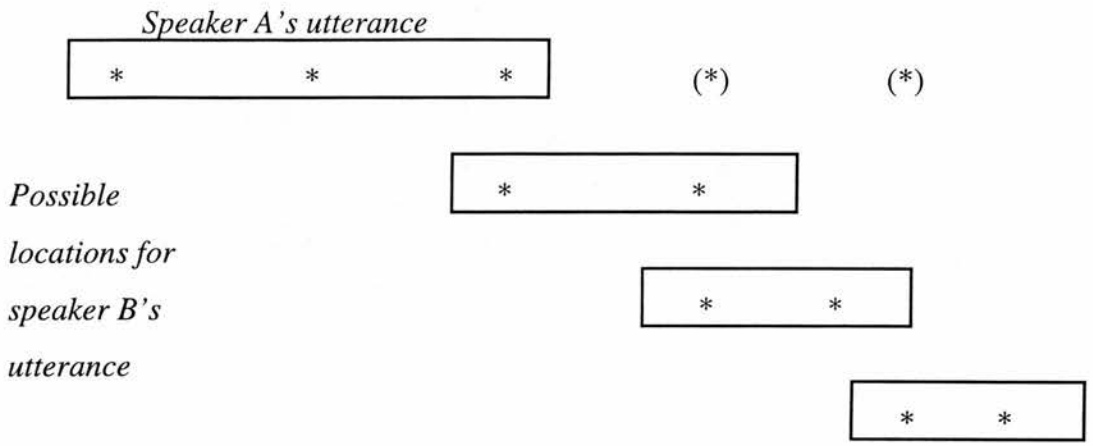
5.4.1 Introduction

The two experiments reported on in section 1.3 above gave no reason to suppose that subjects were able to agree on the durations of between intervals which would make an exchange appear 'natural'. There was therefore no evidence of any application of a perceptual isochrony. To follow the issue we turn to a direct analysis of recorded dialogue, because of possible limitations with the experiments. For example, the fact that the start-intervals had such a bearing on the finish-intervals may have arisen because many of the subjects adopted a conservative strategy in their choices of suitable between-intervals.

The analysis reported on in this section set out to test whether between intervals measured in dialogue taken from the Map Task Corpus fitted into isochronous sequences across turn transitions. It was reasoned that if isochronous sequences form the unmarked case of turn-taking coordination, they should be observable in relatively large numbers in the data. As noted earlier, there are two factors to consider in an analysis of between-intervals and isochrony across turn transitions. The first is the presumption in the rhythmic coordination hypothesis that perceptual isochrony does not occur only when the first stressed syllable of N's utterance falls on the first imaginary beat after the end of C's utterance. The first stressed syllable may occur on the first, second, or third imaginary beat. It may even

coincide with a stressed syllable before C's utterance has finished, as 2) below shows.

2)



The second point is that *actual* analysis of inter-stress intervals is made here, although the claim made by the rhythmic coordination hypothesis is that isochrony is a *perceptual* phenomenon. This discrepancy can however be accounted for by Couper-Kuhlen's observation that a margin of ±20-30% of a previous inter-stress interval is permitted before the perception of an isochronous sequence of stresses breaks down. This is the assumption adopted here, although it is not without its problems (see chapter 2).

The analysis reported on here marked exchanges for prosodic prominences approximating to pitch accents. The intervals between these prominences were measured, both within and across utterances. The theory predicts that in a large number of exchanges the mean inter-stress interval of the first utterance in the exchange (the Within Interval) would be approximately equal to (± a maximum of 30%) the between-interval, or multiples thereof. In fact, the prediction was not supported by the data, even with a relatively loose definition of isochrony.

On the other hand, when between-intervals were analysed according to move type, there were positive results showing that there was a relationship between the two, and that some move types involved longer or shorter between-intervals from others. These were particularly interesting results because they gave preliminary

indications that the context of utterance was a significant factor in the timing of turn-taking.

5.4.1.1 Method

For this analysis, I selected 624 exchanges[19] from quad 4, in the +eye contact condition, of the Map Task Corpus. A roughly equal number of exchanges were selected from each dialogue. All exchanges were relatively fluent. 'Fluent' here refers to exchanges without lengthy delays, hesitations or disfluencies.

Prosodic prominences approximating to pitch accents were marked in these exchanges. In some instances it was possible to test subjective judgements against F0 pitch tracks, but because the perception of stressed syllables depends on more than F0 values, judgements took precedence over F0 measurements.

Of the 624 exchanges marked for pitch accent, only 343 had initial utterances including more than two prominences. Since rhythmic patterns can only be established when there are three or more prominences, exchanges with shorter first utterances could not be used to test the rhythmic hypothesis. Note that exchanges were included when the second utterance in the exchange consisted of only one prominence, since the rhythmic hypothesis requires only that the *first* prominent syllable in a speaker's utterance must coincide with the isochronous chain set up in the previous speaker's utterance. Any subsequent prominent syllables are therefore irrelevant to a test of the hypothesis. Exchanges were *not* selected according to any isochronous or non-isochronous criteria. It was enough that exchanges consisted of a minimal number of prominent syllables.

Within Intervals were calculated from the durations between the vowel onsets of the pitch accented syllables. between-intervals were calculated similarly. The move category of each utterance was noted, so that some preliminary account of the effects of context could be made.

---

[19]These were not selected using the same criteria as outlined in Chapter 4, but were chosen at random.

5.4.1.2 Results

Figure 10 shows the distribution of rhythm ratios, where the dark bars represent on beat exchanges (where ratios are within 0.25 of an integer), and the light bars represent off beat exchanges. No significant difference was found between the number of on-beat exchanges (n = 171, or 49.9% of all exchanges) and off-beat exchanges (n = 172, or 50.1% of all exchanges). The hypothesis only claims, however, that a notional beat set up by a perceptually isochronous sequence degrades after perhaps only three or four beats. That is, without sufficient stimulus, there is a limit to the number of imaginary beats that a listener will be able to follow before 'losing' the rhythm. When the lower rhythm ratios (from 0 to 3.75) were considered separately, it was found that 152 (44.3% of all exchanges) were on beat, whereas 158 (46.1% of all exchanges) were off beat. Even at these lower ratios, where perceptual isochrony ought to be at its most pronounced, there is no evidence that ratios within ±25% of an integer are more common than those which are not.

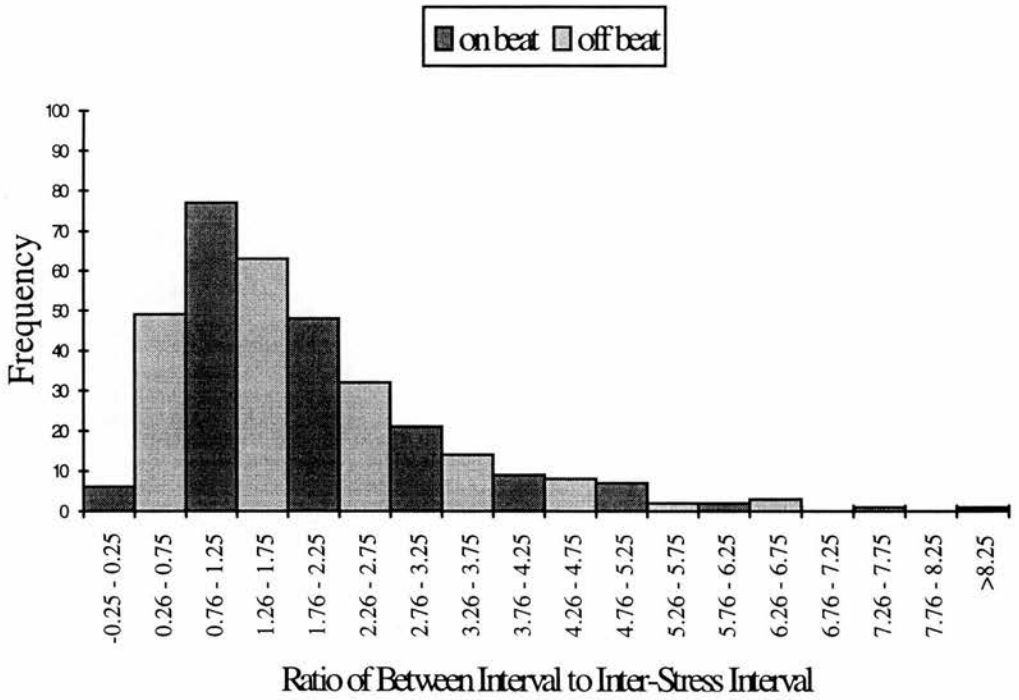


Figure 10 - The ratio of between-intervals to inter-stress intervals

One might argue that a noticeable feature of Figure 10 is the bias of results toward a ratio of approximately 1, and that this might suggest some rhythmic effect. But the

presence of a relatively high proportion (22.4% of all exchanges) of on-beat exchanges in the range 0.76 - 1.25 need not be caused by rhythmic factors. The histogram in Figure 10 fits a skewed distribution, and it may simply be coincidence that the modal value falls between 0.76 and 1.25. Also, 18.4% of exchanges fell in the range of ratios between 1.26 and 1.75 - an off-beat range.

Figure 11 below shows a scatter plot of between-intervals against Within Intervals. There is no significant correlation of the between-intervals against mean inter-stress intervals (R = 0.050, $p = 0.427$). There does appear to be some degree of clustering of intervals below about 1000ms for within intervals, and below about 1800ms for between intervals. But there is very little relationship between the two other than this. Even if there were, one would at least expect a gradient of 1 to the regression line. This was not the case here ($y = 882.9 + 0.0831x$).

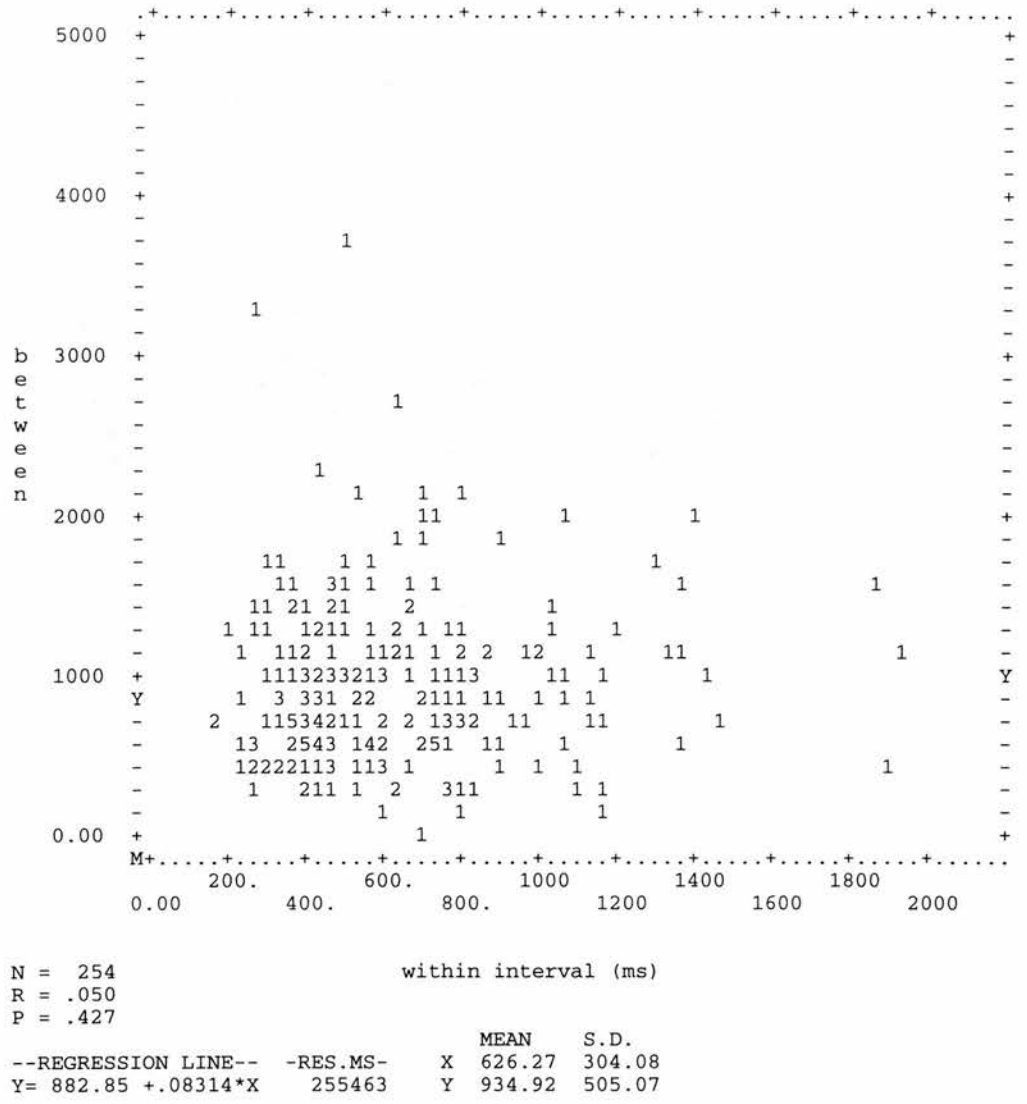


Figure 11 - Plot of the correlation of Within Intervals to between-intervals

Finally, I considered move type. In the Map Task Corpus, each utterance or part of an utterance had been subdivided into twelve categories of move units, coded according to its discourse function (Carletta et al. 1995). I found that exchanges involving different move types had different mean between-intervals. Figure 12 shows the mean between-intervals for each exchange, categorised according to the move type of the *second* move in each exchange. These results are only preliminary, however, because of the small sample of data and thus the low frequency of cases of some of the move types. Nevertheless, Figure 12 below shows differences in the between-intervals when classed into different move categories. The fact that such differences

are relatively easy to observe - and yet results from an analysis of isochrony are uncertain and give no definite positive results - seems to indicate that an analysis based on contextual factors is preferable to one based on rhythmic factors.

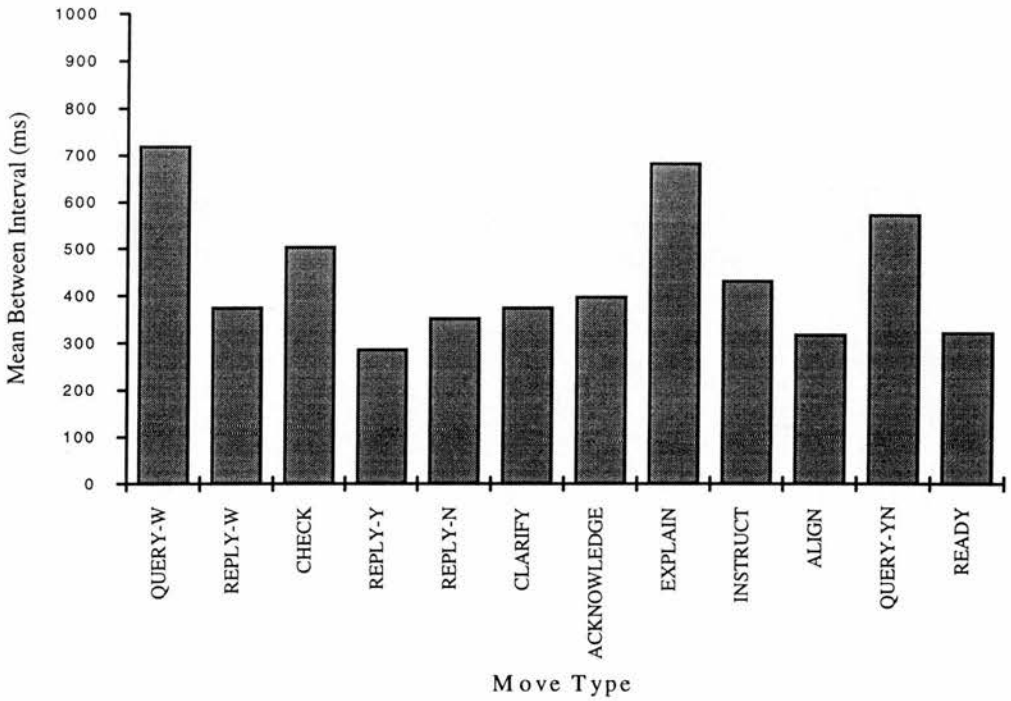


Figure 12 - Histogram of between-intervals according to move type

5.4.1.3 Conclusions and Discussion

The results offer no evidence in favour of the rhythmic hypothesis. between-intervals and Within Intervals did not appear to be correlated to any significant degree. Contrary to expectations, I found no significant difference between the frequency of on-beat exchanges and off-beat exchanges. The hypothesis predicted that on-beat exchanges should have been more common, particularly at lower values of the rhythm ratio.

The observation that the distribution of ratios was skewed toward 1 did not appear to result from any rhythmic effect. Rather, it resulted from a tendency for between-intervals and Within Intervals to cluster within certain ranges which are coincidentally similar.

It could be argued that the rhythmic hypothesis is a perceptual model, and as such detailed instrumental analyses of inter-stress intervals are not necessarily meaningful. There are two problems with this. First, perceptual studies run the risk of being subject to too many contextual variables. There is moreover little support for the rhythmic hypothesis from perceptual studies. In the experiments reported on in sections 1.2 and 1.3, it was found that subjects were unable to decide consistently whether a given between interval was any more or less 'natural' than any other, within certain extremes.

The second problem is that without some form of instrumental support, the perceptually-based rhythmic hypothesis cannot be used meaningfully to generate a predictive model of the timing of turn-taking. It seems reasonable for a certain degree of manoeuvre to be allowed in the analysis, such that Between and Within Intervals would not need to be exactly the same, but could differ by, say, ±25% of each other. Even if this is accounted for, the results in Figure 1 do not cluster around ratios of 1 or 2 sufficiently for strong evidence of a rhythmic effect. And if one were to increase the size of the window, the usefulness of the model is further diminished.

This does not mean that rhythmic structure plays no role in turn-taking. It may be that rhythm aids prediction of TRP location (together with factors such as syntax, pragmatics, intonation, and gesture). It is also quite possible that the tempo of the speech surrounding any given exchange may to some extent be correlated to between intervals. But any role played by rhythm would appear to be small, particularly compared to contextual factors such as move category.

5.5 Analysis of Map-Task Data: automatically-labelled data

5.5.1 Introduction

The analysis reported on in the previous section relied on a selection of only 343 exchanges. There are however over 17,000 possible exchanges available in the Map Task. An analysis using all, or most, of these would therefore offer more reliability. To avoid individually labelling all the exchanges in the Map Task for pitch accents, it was necessary to develop a system which could mark speech for prominent syllables

with as much consistency as a group of expert listeners. The next section describes this system. The model was developed by Matthew Aylett, at the Human Communication Resource Centre, Edinburgh, with cooperation from myself.

5.5.2 Automatic Stress Labelling Model

Whether a syllable is perceived as stressed or not depends on several factors - pitch, duration, amplitude and vowel quality being the more significant. Moreover, the perception of stress is relative, and depends on the degree of stress of surrounding syllables. Ideally, then, a model which is able to determine likely locations of stressed syllables in speech should take all these factors into account. Here, however, we used only duration as a measure of stress, because word duration data was readily available in the Map Task Corpus.

Although stress occurs at the syllabic level, the model used here operated at a word level. That is, for a citation form of a word[20] such as *paper*, the first syllable is prominent, whereas the second syllable is not. Although monosyllables are thought of as having lexical stress, in connected speech many words occur without any prominent syllables. Very often monosyllabic closed class words[21] are not prominent in connected speech, whereas open class words tend to be prominent. Closed class words tend only to be stressed for particular emphasis. For example, consider the difference between the two utterances in 3), where capitals denote a stressed word.

3)

a) Did Marion give it to him?

b) Did Marion give it to HIM?

---

[20]The citation form of a word refers to the way a word is pronounced and stressed if uttered in isolation.

[21]Closed class words belong to lexical categories which have only a finite number of possible members, such as determiners or prepositions. Open class words belong to categories which could potentially have an infinite number of members, such as verbs and nouns.

Further, a stressed syllable may receive primary, secondary, or even tertiary stress, depending on its degree of stress. For the purposes of this model, if any of the syllables in a word were considered to be realised as having primary stress, then the whole word would be considered as stressed.

This view was taken strategically, and is not a theoretical stance. Since the Map Task Corpus offers word durations but not syllable durations, a durational model has to determine stress from the duration of a word. The basic assumption behind the model was that as the actual duration of a word drops relative to the predicted duration of that word in a stressed form, so it becomes increasingly unlikely that the word is stressed. For example, suppose that the expected duration of the closed class word *him* were calculated to be 150ms. The theory maintains that as the duration of a token of *him* (as measured in real speech) drops below 150ms, so it becomes increasingly likely that that token is not a stressed token of *him*.

The expected duration of a word can be calculated if the mean durations and standard deviations of each segment in a word are known. Expected durations of the citation forms of words were calculated using two sources of information. Data from the CELEX database (see Burnage, 1990) was used to provide the basic phonemic structure of the words, and their stress patterns. The ATR database (see for example Campbell, 1993) and a phonetically coded Map Task dialogue were used to provide information on mean segment durations.

Five models were tested, each one using a different method for calculating mean segment duration. The theory behind each one is listed below.

i) Simple model

All phonemes were assumed to have the same duration, and the same mean log duration was used. This 'standard' duration was taken from the ATR database. In effect, there was no differentiation between phonemes. Expected word durations were therefore calculated by multiplying the number of segments in the citation form of a word by the single segment duration.

ii) Simple model with post-modifiers

This was the same model as i), only with consideration of syllabic information. For example, the expected durations of segments were altered according to the number of segments in that syllable, such that the expected duration was decreased as the number of segments in a syllable increased. Also, account was taken of whether a syllable received primary stress in a citation form, whether the syllable was initial, mid, or final, whether it was the only syllable in the word, and whether the syllable was phrase initial or phrase final. The extent to which expected durations of segments were altered was based on data generated by Campbell (1993) using the ATR database.

iii) Syllabic context distribution

This model was the same as model ii, only pre-labelled Map Task Corpus phonetic and stress data were used. This ensured that the test data was of the same basic form as the data that the model would eventually be run on (namely, on dialogues where speakers had Scottish accents).

iv) Phonemic distribution

Mean log distributions were calculated for each phoneme separately using data from the ATR database, and expected word durations calculated by summing the means. A problem arose because of the discrepancy between the Scottish accents which occur in the Map Task Corpus, and the Standard English phonemic representations used in the CELEX database. It was decided, for the purposes of this model, that differences in pronunciation would be ignored, and that therefore only rough approximations of expected word durations were produced. This decision was made largely for practical reasons, and because no reliable data was to hand which provided both information on Scottish phonology and information on observed segment durations in Scottish English.

v) Phonemic distributions and syllabic context post-modifiers

This model was similar to model ii, but now the phonemic make-up of each word was not ignored, and was based on the same criteria as model iv.

5.5.2.1 Determination of Stress

Each of the models outlined above was used to produce expected segment durations and standard deviations. A way was therefore needed to calculate predicted word duration, and thereby obtain an estimation of whether a word was stressed or unstressed.

The predicted length, d_p, of any word may be expressed as:

4)

$$d_p = \sum_{i=1}^{n} exp\,(\mu_i + k\sigma_i)$$

where n = the number of phonemes in a word,
k = a constant function of average segment length,
μ = the mean log duration of a segment, and
σ = the standard deviation of the distribution of segment duration.
The assumption is that the distribution of log segment durations is normal.

The difference between the duration of a segment predicted by one of the models and its observed duration may be expressed in terms of the value of k in the above equation - the k-score. This is the number of standard deviations each actual segment duration is from the expected log mean duration. It is therefore effectively a measure of how much the predicted length of a segment has to be 'squashed' or 'stretched' to fit the actual length.

The number of phonemes in a word, and the mean log duration and standard deviation of each segment in the word were known (from data in the ATR database and one phonetically labelled dialogue in the Map Task Corpus). The only unknown was therefore the k-score. It is important to note that the k-score for a word was calculated on the assumption that if a word's duration were shorter or longer than the expected duration, all segments within that word were squashed or stretched equally.

K-scores were calculated by assuming an initial k-score of 0 for each segment in a word. If the resulting value for the predicted word duration (according to the equation above) was higher than the observed word duration, a lower k-score (-0.001) was used. If the predicted word duration was lower than the observed duration, a higher k-score (0.001) was used. This process was continued until the predicted and actual word durations were the same. The value of the k-score at this point was taken as a measure of the difference between predicted and observed word durations.

K-scores were used to determine whether a word was to be counted as stressed or unstressed. If the k-score for any word was below a certain threshold, this meant that the word was shorter than the model predicted for a stressed word in that context, and therefore that the word was not stressed. If the k-score was above the threshold value, the word was considered to be stressed. Note that a different threshold was used for each part of speech. For example, a determiner such as the word *a* is normally unstressed, and the difference in duration between a stressed and unstressed occurrence would be relatively large compared to a word like *gold.*

The level of the threshold was set for each part of speech on the basis of one dialogue in the Map Task Corpus which had previously been manually labelled for stress. The thresholds were calculated by initially setting an arbitrary threshold value, and assigning stress to all the words in a test dialogue. The k-scores for each part of speech were recorded, as well as the numbers of correct and wrong assignments that the model made (a correct assignment was where a stress was assigned to a word which was marked as stressed in the manually-labelled data). The final threshold value was set at a level where the number of incorrectly assigned stresses was equal to approximately 40% of the number of correctly assigned stresses.

5.5.2.2 Method for manually-labelling the test dialogues

Two test dialogues were selected from the Map Task Corpus. Three subjects who were experienced in linguistics were presented with the dialogues on Entropic Research Laboratory's *xwaves* speech software, so that they could see a speech waveform, and hear selected segments of speech as much as they felt necessary. However, subjects were encouraged to make decisions as quickly as possible.

The subjects were asked to decide for each word in the dialogues whether that word sounded prominent in any way. They were not asked to make specific judgements about stress. If a word was perceived as stressed, the subjects were asked to mark the most prominent syllable in that word.[22]All word and syllable boundaries had previously been marked for the subjects.

5.5.2.3 Results of a comparison of different models

Each model was run on a test dialogue, and evaluated in terms of the numbers of stresses which agreed with or differed from the stresses marked in the manually-labelled test dialogue. The results were as shown in Table 1.

Table 1

	correct	*missing*	*extra*
Model 1	589	148	100
Model 2	613	124	114
Model 3	619	118	116
Model 4	636	101	122
Model 5	623	114	112

There was therefore relatively little difference between the performance of the models, except for Model 1 (which used only one standard phoneme duration, and did not take into account any syllabic information). Models 4 and 5 gave marginally the best results. Both these models made use of data from the ATR database. The problem with these models, therefore, was that they used segmental data from one corpus on a test dialogue from a different (Map Task) corpus. The accents of English used in the two corpuses differed, and so the reliability of the results of Models 4 and 5 had to be called into question. Therefore, although Model 3 did not perform quite

---

[22]The marking was done with the xlabel software package, which stored the exact location of each label within a dialogue.

as well, it was chosen as the model to be applied to all the data in the Map Task Corpus because it was more theoretically justifiable.

A final evaluation using Model 3, when compared to test dialogues labelled by three subjects was as shown in Table 2 below.

Table 2

	correct	*missing*	*extra*
a-A	399	176	182
b-A	389	115	192
c-A	253	63	328
b-a	390	114	185
c-a	260	56	315
c-b	262	54	242

Where A is Model 3, and a, b, and c are the subjects.

There is good agreement between a-A, b-A, and b-a. In other words, the model predicted stress placement much like two of the subjects. The third subject seemed to agree equally poorly with the other subjects as with the model.

Therefore, Model 3 was chosen, and the results here indicate that it was able to perform as well as two subjects in predicting stress placement.

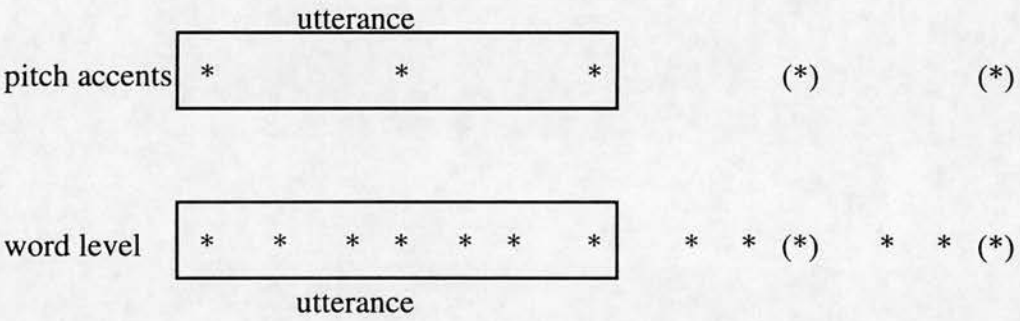
5.5.3 Method

The model outlined in section 1.5.2 above was applied to the set of Map Task exchanges which had not been filtered out by the elimination processes mentioned in Chapter 4. It will be recalled that these processes filtered out exchanges where one utterance could not reasonably be considered a response to an immediately preceding or overlapping utterance.

The model marked each *word* as either stressed or unstressed. This meant that while pitch accents were not marked directly, they were marked indirectly in some of the stressed words. The rhythmic coordination operates with what approximates as a pitch accent level of stress (although see Chapter 2 for a discussion of the problems

in deciding on an appropriate level of stress). However, because of the nature of the rhythmic coordination hypothesis, it was assumed that if the first stressed syllable (or pitch accent) in an utterance coincided with the beat set up by a perceived isochrony in a previous utterance, the same should be true of stresses at a word level. Basically, although the mean inter-stress intervals within an utterance will be different according to whether word-level stresses or pitch accents are used, the imaginary beats will still correspond to some multiple of the mean inter-stress interval (although the multiple will be higher than the factor of 1, 2 or 3 assumed for a pitch-accented system of rhythmic coordination). Example 5) below demonstrates this, where an asterisk represents a realised stress, and an asterisk in parentheses represents an imaginary beat.

5)



Once the model had been applied, further cases had to be eliminated. For example, a minimal condition for perceptual isochrony is that there must be *at least* three beats for any rhythmic structure to be built up. Therefore, if any utterance in the first part of an exchange consisted of less than three stressed words, then the whole exchange was eliminated from the analysis. This second process of elimination left 3104 exchanges.

If the rhythmic coordination hypothesis is correct, in the majority of exchanges the first stressed syllable in the second utterance in an exchange should be aligned with an imaginary beat set up by the perceived isochrony of the first utterance in an exchange. In other words, the between-interval should be equal to the mean Within Interval, or some multiple of it. As a test of the hypothesis, the between-

interval was therefore divided by the mean inter-stress interval for each of the 3104 exchanges, to give a *rhythm ratio*. This ratio should be approximately equal to an integer. Negative integers may be obtained - these occur when there is overlap between two utterances, and the between-interval is negative. A ratio of zero occurs when the between-interval is 0ms in duration (when there is latching). As mentioned earlier, there is a permissible margin of error before the perception of isochrony is supposed to disappear. Couper-Kuhlen (1993) found the level to be somewhere between 20% and 30% of a previous inter-stress interval. A 25% margin of error has been used here. Any exchange where the rhythm ratio was equal to an integer ±25% is termed *on beat*. Any exchange where the rhythm ratio falls outwith the range of an integer ±25% is termed *off beat*.

5.5.4 Results

Figure 13 does not show any major differences between on beat exchanges (marked by the dark bars) and off beat exchanges (marked by the light bars). Of the 3104 exchanges, 1557 (50.2%) were on beat, whereas 1546 (49.8%) were off beat.

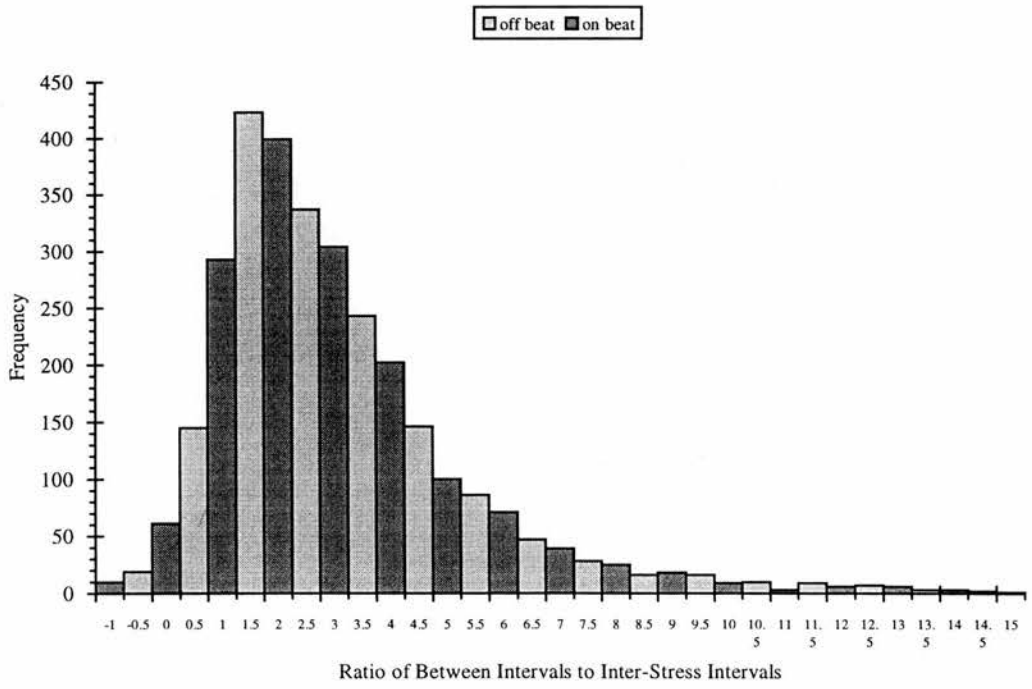


Figure 13 - The ratio of between intervals to inter-stress intervals. The centre of each bin is marked on the histogram, although each bin runs from +0.25 to -0.25 of the centre

If the rhythmic coordination hypothesis were correct, it would be expected that there would be significantly more on-beat exchanges than off-beat exchanges, particularly with lower integers (rhythmic coordination is theoretically more likely at lower multiples of the mean inter-stress interval because of memory limitations of participants in a conversation). However, if lower ratio values (from -1.25 to 4.75) are used, there are 1269 (40.9%) on beat exchanges and 1313 (42.3%) off beat exchanges.

A regression of inter-stress intervals and between-intervals was carried out. This did not reveal any significant correlation between the two variables.

5.5.5 Conclusions

The analysis reported here found no instrumental evidence in favour of the rhythmic coordination hypothesis. The prediction from this hypothesis is that the majority of exchanges will consist of stressed syllables which are arranged in perceptually isochronous sequences across turns. Because they are perceptually isochronous, the sequences of stressed syllables will not have exactly equal durations between them. There has to be some degree of freedom within which perceptual isochrony holds, and this was taken into account in this analysis. For isochrony to hold across turns under this basis, the ratio of a between-interval to an inter-stress interval should be equal to an integer, ±25%. No evidence was found that this was true for the majority of exchanges.

Other evidence does not favour the rhythmic coordination hypothesis. Of the 11,017 exchanges that were available for analysis, only 3104 (28.2%) exchanges could be used for a rhythmic analysis. This was because of the constraint which meant that there had to be at least three stressed syllables in the first utterance in an exchange. The fact that such a relatively low percentage of exchanges could be considered suggests that perceptual isochrony must have a minor role in the coordination of turn-taking, if it is used at all. It means also that rather than 50.2% of exchanges being on-beat (using the total of 3104 exchanges), only 14.1% of all possible exchanges (using a total of 11,017 exchanges) could be described as on-beat.

The rest were either potentially isochronous (but were not found to be isochronous), or could not even potentially be considered to be isochronous.

This evidence does not prove that perceptual isochrony is never used in the coordination of turn-taking, since on-beat exchanges do exist. However, the question here is whether they exist in significant numbers that their status as the default case of coordination is justified. There is no evidence that this is the case.

5.6 Summary

This chapter has reported three experiments and two sets of data analysis which were carried out to test the rhythmic coordination hypothesis. None of these provided conclusive evidence in favour of the hypothesis. I carried out the perceptual experiments to test the notion that under the rhythmic hypothesis subjects should have been able to differentiate between preferred and dispreferred turn transitions, and that the preferred transitions should have consisted of isochronous sequences (perceptually isochronous sequences form the unmarked case of turn transition). However, subjects were only able to make consistent judgements on the basis of the between-interval that they were first presented with, rather than any rhythmic basis.

The two data analyses - using exchanges that were manually and automatically labelled for stress - did not provide any evidence for the rhythmic hypothesis. These analyses tested the prediction that in the majority of cases there should be a perceptually isochronous sequence across a turn transition, where a certain margin of freedom was permitted in the duration of a between-interval (±25% of the mean inter-stress interval) before the perception of isochrony was considered to be lost. However, no significant difference was found in the numbers of exchanges which fell into an isochronous pattern, and those which did not. Although this does not disprove the role of isochrony in the coordination of turn-taking, if it were to be used as a major coordinating factor then it would be expected that the proportion of exchanges which had an isochronous sequence would be much greater.

6. An Analysis of the Influence of Context of Utterance on Inter-Speaker Intervals

6.1 Introduction

Two factors form the basis of the timing of turn-taking - *intelligibility* and *earliest possible start* (Sacks et al., 1974). The intelligibility constraint requires that a coordinated exchange of turns should be understandable, and that there should not be so much overlap that neither speaker understands the other. The constraint of earliest possible start counterbalances this, because it dictates that N should start as soon as possible after C has finished speaking. In effect, the conversational floor is a scarce resource which has to be managed as efficiently as possible. But as Couper-Kuhlen (1993) argues, if only these two constraints were applied we should always find latching in turn-taking (latching is the situation where there is an inter-speaker interval of 0ms). Clearly, not all exchanges are latched. Therefore, some other factor or factors must act alongside the intelligibility and earliest possible start constraints.

The previous chapter presented experiments and analyses which tested the hypothesis that perceived isochrony is this other factor in the coordination of turn-taking. I found no evidence to support this hypothesis. An alternative hypothesis is that consideration must be given to the limitations of perception and planning time (the 'cognitive' context), and the need in conversation to communicate information on several levels (the 'communicative' context. Also see the theories of Clark (1996), and Chapter 2 of this thesis). The latter constraint may consist of the communication of a range of information, such as social status or common ground, and may alter the degree of overlap or interval in the temporal coordination of turn-taking which is acceptable to the participants, because the inter-speaker interval acts to signal factors such as their relationship, whether they understand each other, or what common ground may hold between them.

The variables which define the communicative context of utterance - the *communicative variables* - may be thought of as acting globally in a conversation, because they will not change during the course of a single conversation.[23] They determine the expectations that each speaker has about the other, and therefore the expectations that both have about the general ranges of inter-speaker intervals that would be tolerated. As the communicative context of a conversation changes, so the 'window' of acceptable inter-speaker intervals should shift. The assumption here is that both participants are similarly aware of the communicative context, and therefore of the acceptable window size. They must each also be aware that the other participant is similarly aware. The exact relationship between inter-speaker interval and communicative context is uncertain, and as was noted in Chapter 2, I am unaware of any empirical research into this area.

Some communicative variables act locally. Participants may agree or disagree about some matter, or either may feel the need to put their respective points across at the expense of the other's. Variables such as level of agreement, disagreement, and topic could vary from one exchange to the next. Participants might use their (presumed) mutual and implicit knowledge of a link between inter-speaker interval duration and this social information to signal agreement/disagreement through interruptions, backchannelling, or smooth transitions.

The cognitive context of an exchange is related to the amount of processing that may be required to plan an utterance and make a response. The general assumption here is that the greater the complexity of an utterance, the more comprehension and/or planning time will be required (cf. *The Principle of Processing Time* - Clark, 1996, and work on chronometrics). Cognitive factors apply locally, because the amount of processing required by participants varies from one utterance to another and from one exchange to another. As with social variables, both participants are presumed to be similarly and mutually aware of the cognitive

---

[23]Or if they do, the general parameters of the conversation will change, and one might claim that a new conversation has been started.

context, and to have an implicit understanding of the relationship between the context and inter-speaker intervals.

It is important to stress that the cognitive and communicative contexts cannot readily be separated from one another at anything more than an abstract level. The planning time involved in turn-taking may be a function of the communicative features participants are attempting to convey, and vice versa. For example, in this analysis I found it difficult to categorise a variable such as eye contact as being purely 'cognitive' or purely 'communicative'. No such gross categorisation was used, therefore.

A further element is the predictability of turn closure. It was shown in chapter 2 how N must rely on a means of predicting the end of C's turn. Although it is not certain exactly how conversationalists do this, prosodic, syntactic, pragmatic and gestural cues play roles (Ford & Thompson, 1995). This research assumes some capacity to predict TRPs, which is applied similarly by different speakers in different contexts.

A test of the linear timing hypothesis, therefore, is to analyse a series of exchanges and to determine what relationship there is between the inter-speaker interval for each one and the social and cognitive characteristics of that exchange. Some factors - notably the categories of move and game type involved in an exchange - can account for differences in inter-speaker intervals in a systematic way. The general approach was therefore to use a ready-made, ready-coded corpus of task-oriented dialogue (the HCRC Map Task Corpus). Each type of code was, where possible, considered for the ways that it might affect inter-speaker intervals. Finally, the effects were analysed. The results support the hypothesis that planning and decision time, with the basic constraints of earliest possible start and intelligibility, play a major role in the timing and coordination of turn-taking.

6.2 Variables

6.2.1 Familiarity

Familiarity was labelled in the Map Task as a binary variable. Naturally, familiarity with another person is a matter of degree, and such a gross classification of a notion of 'familiarity' will always involve difficulties. It was not certain what relationship there would be between a possible familiarity effect and mean inter-speaker interval. Familiar participants may tolerate shorter mean intervals because they feel free to overlap one another's speech. On the other hand, familiar participants may be more tolerant of longer mean intervals because they are feel under no social pressure to prevent 'awkwardly long' silences.

6.2.2 Sex

Sex differences were also accounted for in the corpus. This variable was included in this analysis because it was thought that there might be differences in mean inter-speaker intervals depending on whether the two participants were different-sex, both male, or both female, because of possible socially-determined rules governing the interaction of mixed- and same-sex groups. Sex may play a role in determining inter-speaker interval duration, particularly in mixed-sex interactions.

Some caution was exercised with the Sex variable. There was an uneven distribution of sex differences in the familiar dialogues in the Map Task Corpus, as shown in Figure 1.

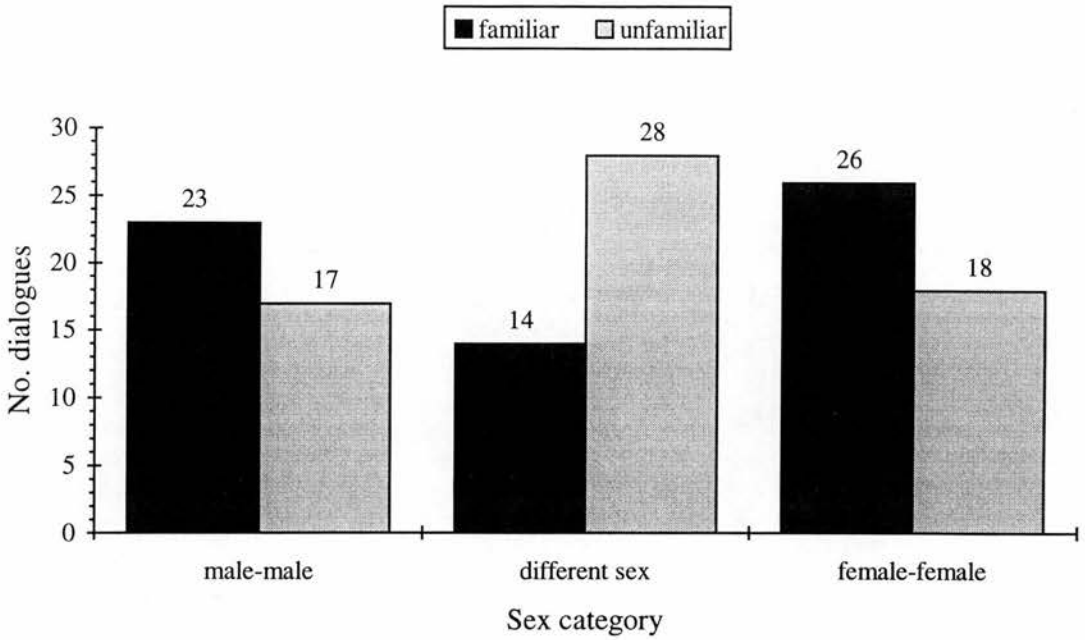


Figure 1 - Number of dialogues split according to sex category. (n = 126)

Of the 63 familiar dialogues, 49 (77.8%) involved pairs that were either all-male or all-female. Only 35 out of the 63 unfamiliar dialogues (55.6%) involved same-sex pairs of speakers.

6.2.3 Role

In each dialogue in the Map Task each participant was assigned the role of either *instruction giver* or *instruction follower*. In the Map Task dialogues this variable corresponded only to the presence or absence of a pre-printed route on the speaker's map. Since role is assigned for the duration of a dialogue and since the status is ultimately crossed with speaker, the Role variable is treated as a globally applicable variable. But in reality, the roles of the two participants may vary slightly throughout a Map Task dialogue. For example, an information follower may very often give instructions to an information giver.

I supposed that there would be a difference in inter-speaker interval duration according to whether there was a switch from giver to follower or vice versa, because of the different planning and decision requirements of each speaker. An instruction giver may be required to plan ahead more than an instruction follower, and to

develop strategies for conveying information as efficiently as possible. This extra planning would require more time, and consequently inter-speaker intervals preceding a giver's utterance may be longer than those preceding a follower's utterance.

Differences may also arise here through association with other factors. For example, givers and followers might depend on different categories of move (move ccategory, as explained below, is considered a cognitive variable). A giver may be expected to give more instructions than a follower, and to start more new conversational games. Givers also have the opportunity to plan ahead more than followers, because it is they who have the route, and it is they who can dictate the general strategy.

6.2.4 Eyecontact

This variable is a measure of whether the two participants were *able* to see one another, and not whether they were actually looking at one another at any stage. The corpus was split into eye contact dialogues and no eye contact dialogues. In the no eye contact dialogues, a sheet of card was placed between the two participants so that they could not see one another, but could still hear each other clearly. Eyecontact was therefore a globally applicable variable, and could not vary from one exchange to the next. The mean inter-speaker intervals of eye contact and no eye contact dialogues might be significantly different from one another, although no definite prediction was possible of which would have the bigger mean.

6.2.5 Conversational Game Boundary

The Map Task Corpus has been labelled for the start and end points for each game, as well as information on whether each game was embedded within other games (see Chapter 2). An embedded game is one which starts before a previous game has finished. In 1) game1 and game2 are not embedded. Game3 is embedded, because it starts before game2 has been completed.

1)

$$[_{game1}\ utterance1 \quad utterance2]_{game1} \quad [_{game2}\ utterance3 \quad [_{game3}\ utterance4 \quad utterance5]_{game3}\ ]_{game2}$$

The coding allowed a comparison of inter-speaker intervals within games, and across game boundaries for both embedded and non-embedded games.

If the introduction of a new game depends on some extra decision time or planning time (according to the principle of processing time) then inter-speaker intervals should be longer between utterances separated by a game boundary than between those within a game. So, in example 1) above, the inter-speaker intervals between utterances 1 and 2, and utterances 4 and 5 would be shorter than the intervals between utterances 2 and 3, and utterances 3 and 4. Because there might be added processing each time a new game is started before a previous one is finished, this effect may apply to embedded games, giving longer inter-speaker intervals between utterances separated by an embedded game boundary than between those separated by a non-embedded game boundary. In this case, the inter-speaker interval between utterances 3 and 4 (across an embedded game boundary) would be greater than between utterances 2 and 3 (across a non-embedded game boundary).

A preliminary 1-way ANOVA comparing the three levels of Game Boundary variable (embedded game boundary, non-embedded game boundary, and non-boundary) showed that there was no significant difference between embedded and non-embedded game boundaries. The mean inter-speaker interval at game-internal exchanges was 431ms (SD = 726ms, n = 7927). The mean interval for non-embedded game boundary exchanges was 621ms (SD = 775ms, n = 972), and for embedded game boundary exchanges it was 672ms (SD = 872ms, n = 2118). A Scheffé test revealed a significant difference (at the 0.01 level) between the mean inter-speaker intervals of game-internal exchanges and the mean inter-speaker intervals at game boundaries, but not between mean inter-speaker intervals at embedded and non-embedded game boundaries. For the rest of the analysis, therefore, the Game Boundary variable had only two levels - boundary and non-boundary.

6.2.6 Move Category

The Map Task has been coded for two types of move class. One was a 12-way system; the other was a 3-way system (see Chapter 2, and also Carletta et al., 1995). Move coding offered a ready means of accounting for the illocutionary force of utterances, and their syntax. For example, questions were categorised as either y/n-questions or wh-questions. The fact that some wh-questions expect a prompt answer ('*Have you got diamond mine*?'), and others do not ('*Where's the diamond mine on your map*'[24]) is not coded. Nevertheless, it was possible to make the general assumption that different move categories would reflect differences in the function and content of utterances, and therefore differences in processing time.

6.2.6.1 3-class

This system consisted of *initiating moves, response moves,* and *transitional moves* (see Carletta et al., 1995). Initiating moves (*instruct*, *explain*, *check*, *align*, *query-yn*, *query-w*) set up the expectation of a response, and are often found at the start of a new game. Response moves (*acknowledge*, *reply-y*, *reply-n*, *reply-w*, *clarify*) are used within games after an initiating move. Transitional moves consist solely of the *ready* move, and as Carletta et al. (1995) point out, it is not clear whether this should be counted as a distinct move class or not. They were initially included in this analysis (although, as reported below, they were eventually omitted owing to the relatively small number of cases involved).

This system was therefore loosely based around game coding. The two are not directly comparable, however, because the 3-class system of move coding presumes that certain move types will *always* act as initiators, responses, or transitions. If they do, then like game boundaries, initiating moves should be preceded by longer inter-speaker intervals than response moves. Like game boundaries, initiating moves very often introduce a new discourse goal. Formulating that goal may require more

---

[24]Although it seems possible that when a person finds a wh-question particularly difficult, he or she may reply in the negative faster than to a simpler question, where an answer is known but is not immediately retrievable from memory.

planning and decision time. Also, different combinations of initiating, response and transitional moves should give rise to different intervals. An initiator-response sequence could be considered as the default case of a game-internal exchange, involving minimal potential conversational difficulty. An initiator-initiator exchange would appear to involve some conversational problem or difficulty since an attempt has been made to start a new game immediately after such an attempt is made by the other speaker. Accordingly, exchanges involving initiator-initiator combinations should have larger inter-speaker intervals associated with them than initiator-response combinations.

6.2.6.2 12-class

Twelve categories of moves were used. Only the moves immediately preceding the speaker switch (a-moves) and following the speaker switch (b-moves) were considered, because these were most likely to have a direct influence on inter-speaker intervals (see chapter 4 for a discussion of the problems in determining exactly which moves in an utterance may be thought of as being the most salient).

However, it proved impossible to carry out full analyses involving both a-moves and b-moves with all twelve move types, because empty cells were generated for some combinations of move (see Appendix B1 for lists of cell sizes). The sets of moves used was therefore as follows:

a-move	**b-move**
check	*acknowledge*
explain	*align*
instruct	*check*
reply-y	*explain*
reply-w	*instruct*
query-yn	*ready*
	reply-y
	query-yn
	query-w

Note that *acknowledge* moves were not used in this analysis in the a-move position because of the constraints on backchannelling and responses to backchannel signals (as mentioned in Chapter 4).

Different moves should be preceded and followed by different inter-speaker intervals, according to their general function. Although a complete ordering for the 12 classes of moves is not possible, some predictions can nevertheless be made. For example, *reply-w* b-moves might generally be expected to follow a longer interval than an *acknowledge* b-move would, because the sort of processing required to reply with some sort of information (a *reply-w* usually answers a *query-w* move, which acts as a request for information) would generally be greater than to produce an utterance such as '*mm-hmmm*', which acts simply as an acknowledgement that the listener was paying attention.

Either the a-move class or b-move class could influence the inter-speaker interval. If interval duration represents processing time, the b-move would be expected to have the greater influence on inter-speaker duration, because the second speaker in any exchange has *direct* control over when to start speaking, and hence how long the interval is. However, that speaker's choice of move is constrained by the type of a-move that the other speaker has used (*instruct* moves are commonly followed by *acknowledge* moves, but not by *reply-w* moves).

6.2.7 Map Variables (Match, Route, and Contrast)

As explained in Chapter 3, each map was defined in terms of three variables - *Contrast*, *Match*, and *Route*. Contrast is a binary variable, and a map is +Contrast when there is a contrast in the names of two master features for that map (e.g. *east* lake, and *west* lake). When there is no such contrast, a map is -Contrast. Match is also a binary variable. A map is +Match when the value of contrast of the giver's map is the same as that of the follower's map. There are four different routes.

These three variables should give rise to some differences in mean inter-speaker intervals, because of the planning times required to deal with differences in the maps of giver and follower.

6.2.8 Task Familiarity

Familiarity with the Map Task should play a role in the determination of inter-speaker intervals. In the first two dialogues of each quad, no participant had encountered the Map Task before, and would therefore not have developed any strategies for completing the task as efficiently and effectively as possible. Without such strategies, it is likely that more planning will be required, and mean inter-speaker intervals will be greater, than when participants are familiar with the general task.

6.2.9 Shared Landmarks

There may also be differences in the mean inter-speaker intervals of exchanges where first mention was made of shared and unshared landmarks. There may be differences in the amount of decision time required by a next speaker when a new and unshared landmark is introduced by the current speaker, and that this may be reflected in mean inter-speaker intervals.

6.2.10 Deviation Score

Subjects' success at the Map Task had been calculated when the corpus was being coded, by comparing the route drawn on the follower's map with the route printed on the giver's map. The deviation between the two was calculated by placing a grid, marked into square centimetres, over the two routes. Whenever the follower's route differed from the giver's, the difference was measured in the number of squares. All the differences were added for the whole dialogue.

Because deviation score rises as printed and drawn routes diverge, it measures the degree of difficulty that the two participants experienced in completing the task in each dialogue. Although deviation score could have been calculated for parts of a map, and hence parts of a dialogue, it had only been calculated for an entire dialogue, and therefore was not a locally-applicable variable as such. It was nevertheless presumed that changes in deviation score would be caused by local effects in each dialogue.

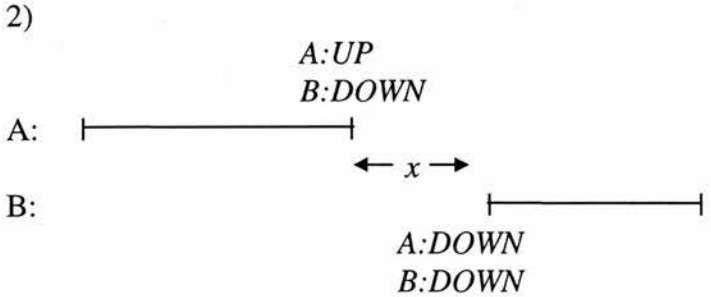
6.2.11 Gaze

In addition to the Eyecontact variable, the Map Task had been coded to include data on which of the two speakers was looking up or down at any point (to the nearest word boundary) in each dialogue. Therefore, the gaze variable was a measure of actual rather than potential eye contact. Gaze at speaker transitions was analysed here, to determine whether it was related to inter-speaker interval duration.

Gaze is marked in the map task corpus in quads 3 and 4 (both eye contact and no eye contact) and quads 7 and 8 (eye contact only). I categorised the data into the following groups: *speaker looks up*, *listener looks up*, *both look up*, *both look down*.

A note needs to be made here of the method by which inter-speaker intervals were assigned to one of the four gaze categories. I measured the gaze status at the start of the first move made by a speaker after a speaker switch, thereby giving an impression of who was looking up or down each time a new utterance was started. I could equally have noted gaze status at the *end* of each utterance, giving information about who was looking up or down each time a speaker finished an utterance. The distinction is important, because it can result in inter-speaker intervals being assigned to different gaze categories. Example 2) below shows that at the end of A's utterance, A is looking up and B is looking down. However, by the start of B's utterance A is looking down. Should the inter-speaker interval (x) be categorised as "*speaker looks up*", or "*both look down*"?

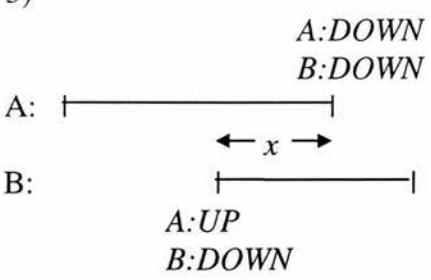
2)



Similar problems hold for cases of overlap. In 3) below, when B starts speaking, A is looking up and B is looking down. But by the time A has finished, both A and B are looking down. Again, should the inter-speaker interval be classed as "*speaker looks*

up" or "*both look* down"? In this case, the situation is more complex, because gaze may have altered more than once during the overlapped speech.

3)



It seems unnecessary to consider gaze *after* B has started speaking if what is required is some description of how gaze may affect inter-speaker intervals. I chose to use the gaze status at the start of an utterance because I thought that the gaze of the two participants was most relevant at the point when a speaker started, rather than when a speaker stopped.

6.2.12 Speaker

Individual differences would almost certainly have an effect on inter-speaker interval durations. Speaker differences seem to encompass all those differences in inter-speaker intervals which cannot readily be explained by any of the variables mentioned above. This variable has relatively little to contribute to an understanding of the factors affecting the timing of turn-taking, other than potentially to say that speaker differences do or do not exist.

6.2.13 Backchannelling

An important consideration in this analysis was whether there is any significant distinction between the timing involved in the backchannel and main channel of a conversation. Backchannelling has primarily social functions. It acts as a form of conversational 'social glue' which indicates that a listener is paying attention to a speaker, forming a sense of social cohesiveness between the two. However, the sorts of utterance that typify backchannelling (for example, utterances beginning with, or

consisting entirely of, *acknowledge* moves) would also be expected to require minimal planning to produce, since they are short, simple utterances with little content. Mean inter-speaker intervals between main channel utterances and backchannelled utterances ought to be shorter than between two main channel utterances.

If significant differences in mean inter-speaker intervals were found then not only would there be evidence for a completely separate status of backchannelling, but also any differences in inter-speaker interval could be the result of differences in the planning time required to produce backchannel signals. An absence of a difference in the distributions of inter-speaker intervals for backchannel and main channel utterances would not counter evidence for the validity of the treatment of backchannelling as a distinct form of communication outwith the main channels of conversation. But it would be increasingly difficult to suggest anything other than that, structurally, backchannelling is treated by interlocutors as if it were in the main channel, but that functionally it may still remain in a separate class because of its social elements. In this analysis, the data was split into backchannel utterances (*acknowledge* moves, and moves with the features 'mumbl' and 'repo'), and main channel utterances. The first speaker's utterance in an exchange was called the *a-turn channel.* The second speaker's utterance was called the *b-turn channel.* More or less by definition, a backchannel is the second part of the exchange. It does not take the conversational floor, so the other speaker need no reply to it. Therefore, we can assume a two-way distinction: main-main, main-backchannel.

6.3 Method

I used the same set of 11,017 exchanges from the HCRC Map Task Corpus as the data for this analysis, as I used in the rhythmic analysis reported on in the previous chapter. The corpus provided data for each of the variables described above. For each exchange in the data set I extracted the inter-speaker interval (the dependent variable) and the values of each of the (independent) variable described above, where

applicable.[25] This allowed me to group inter-speaker intervals into the different levels of each of the independent variables and to find the mean value for that level (for example, the mean inter-speaker interval for all +Eyecontact exchanges). I used the ANOVA test to determine whether there were significant differences between the mean inter-speaker intervals of each level of each variable. If there were significant differences in the levels of an independent variable *x*, this would indicate in turn that *x* plays a significant role in the timing of turn-taking.

Ideally, all the independent variables should be tested for significance together, since this gives the most accurate representation of which independent variables are truly significantly related to the dependent variable. Often an independent variable appears to be significantly related to the dependent variable because of interactions with other independent variables. Running an ANOVA on each independent variable separately would not show this.

However, I could not run an ANOVA on all the levels of all the independent variables simultaneously, because doing this (or even crossing a reduced set of variables) produced many small or empty cells (as Appendix B3 documents). Instead, I was forced to carry out a series of n-way ANOVAs, where each ANOVA crossed a sub-set of all the independent variables.

The first ANOVA in the series used several variables which I thought might play a major role in the determination of inter-speaker interval. Successive ANOVAs crossed new variables with the significant variables from previous ANOVAs, until I had, as far as possible, reduced the complete set of independent variables to a group of variables which were significantly linked to variations in inter-speaker interval durations. I used *post hoc* tests on those significant variables which had more than two levels to show which levels were responsible for the significant effects.

---

[25]For example, some of the exchanges could not be coded for the Gaze variable.

6.4 Results

Table 1 - Summary of the levels of each variable tested using ANOVA analyses.Numbers in body of table represent the number of levels of each variable in that ANOVA.

Section ANOVA appears in / Variable	6.4.1	6.4.2	6.4.3	6.4.4	6.4.5	6.4.5	6.4.5	6.4.6	6.4.7	6.4.8	6.4.9	6.4.9
Gameboun.	2+	2+	2+								2+	
Eyecontact	2+	2+	2+	2+							2+	
Role	2+	2+	2+	2+							2+	
Fam.	2-											
Sex	3-											
Match		2+	2+									
Route		4-										
Contrast		2-										
Taskfam			2+								2+	
A-move				2-	6+	11+						
B-move				2+	9+		12+					
Gaze								8-				
Shrd lndmrk									2-			
Devn score										8+		
A-backch											2-	2-
B-backch											2-	2+

Because of the large number of variables involved in this analysis, only some variables could be compared within a single ANOVA before large numbers of empty or near-empty cells resulted. As a solution to this problem, I chose to compare only some variables in a first ANOVA, and then to cross those variables found to be significant with new independent variables in successive ANOVAs. In this way, each variable could be crossed with at least some of the other variables.

6.4.1 5-way ANOVA - Game Boundary x Eyecontact x Role x Familiarity x Sex.

Game Boundary - 2 levels. Boundary, non-boundary.

Eyecontact - 2 levels. Potential for eye contact, no potential for eye contact.

Role - 2 levels. Giver-follower exchanges, follower-giver exchanges.

Familiarity - 2 levels. Familiar, unfamiliar.

Sex - 3 levels. Both male, both female, one-male-one-female.

This ANOVA showed the following main effects:

6.4.1.1 Role (2 levels)

The 5-way ANOVA crossing the independent variables Game Boundary, Eyecontact, Role, Familiarity, and Sex showed that the Role variable was significant (F (1, 10969) = 30.11, $p < 0.0001$). The mean inter-speaker interval was significantly greater when it fell across a giver-follower speaker switch (mean = 541.4 ms, sd = 824.4 ms, n = 6798) than when it fell across a follower-giver switch (417.6 ms, sd = 658.8 ms, n = 4219). In other words, on average followers were slower to respond to utterances by givers, than vice versa. This was contrary to what was expected with regard to planning time, since one might expect that it is the givers who require more planning time because they have to formulate strategies by which to guide the followers.

A further analysis of the distributions of giver-follower and follower-giver type exchanges reveals that giver-follower switches have higher mean inter-speaker intervals than follower-giver switches because giver-follower switches have a lower proportion of negative intervals (giver-follower = 16.6% overlap; follower-giver = 19.3% overlap). Figures 2 and 3 below give the distributions of inter-speaker intervals associated with giver-follower switches and follower-giver switches.

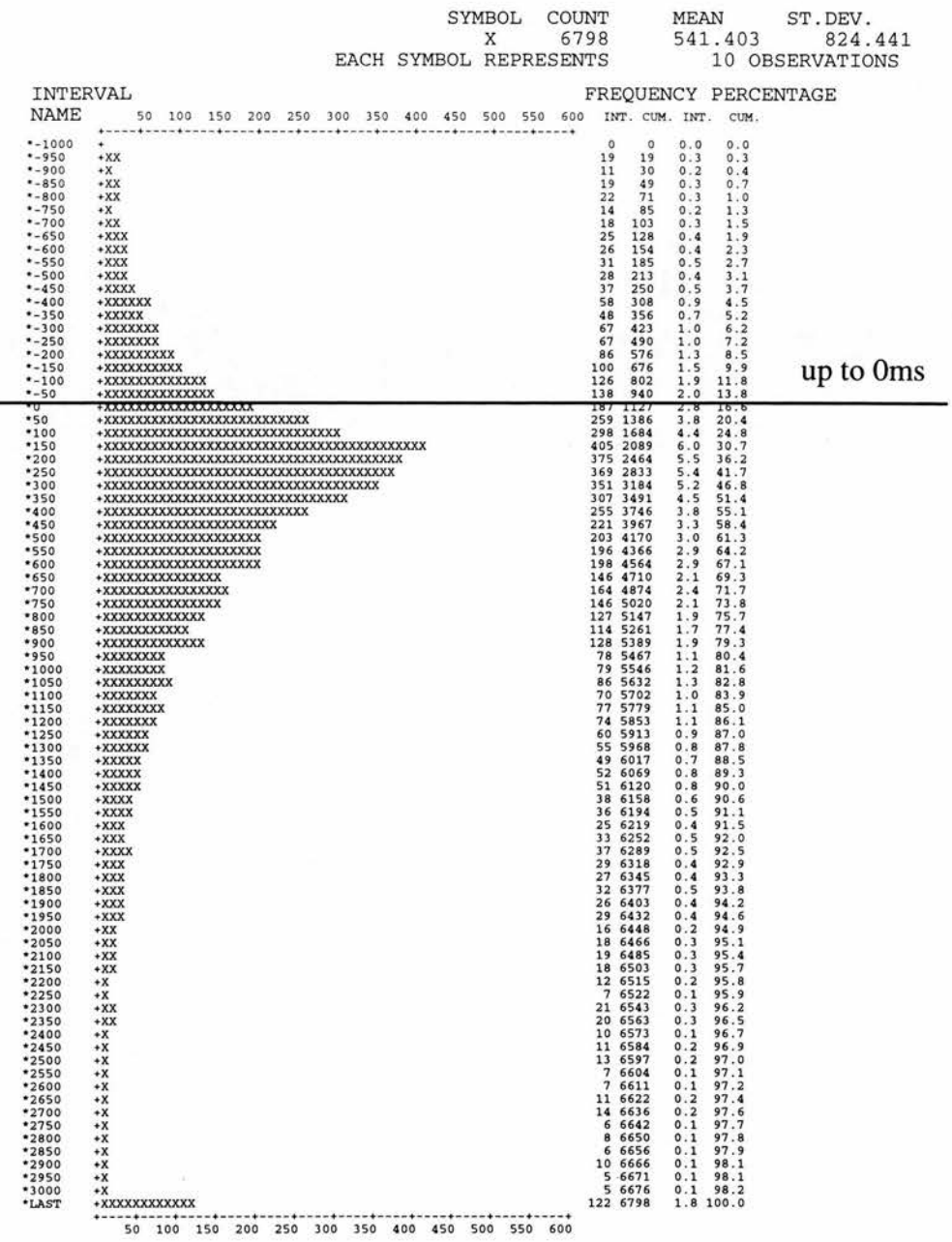


Figure 2 - Distribution of inter-speaker intervals for giver-follower exchanges. Figures on the left indicate upper ranges of the bins

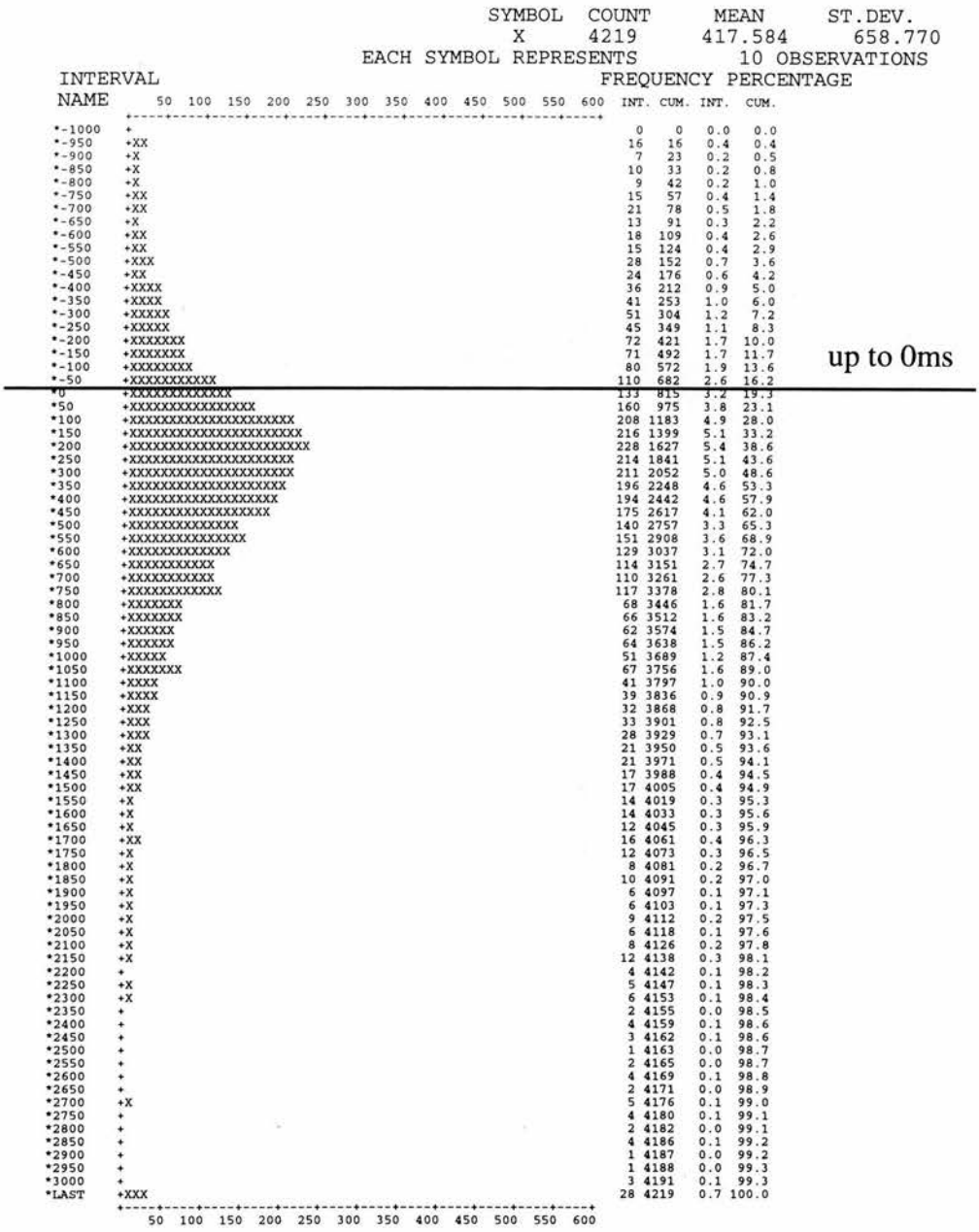


Figure 3 - Distribution of inter-speaker intervals for follower-giver exchanges. Figures on the left indicate upper ranges of the bins

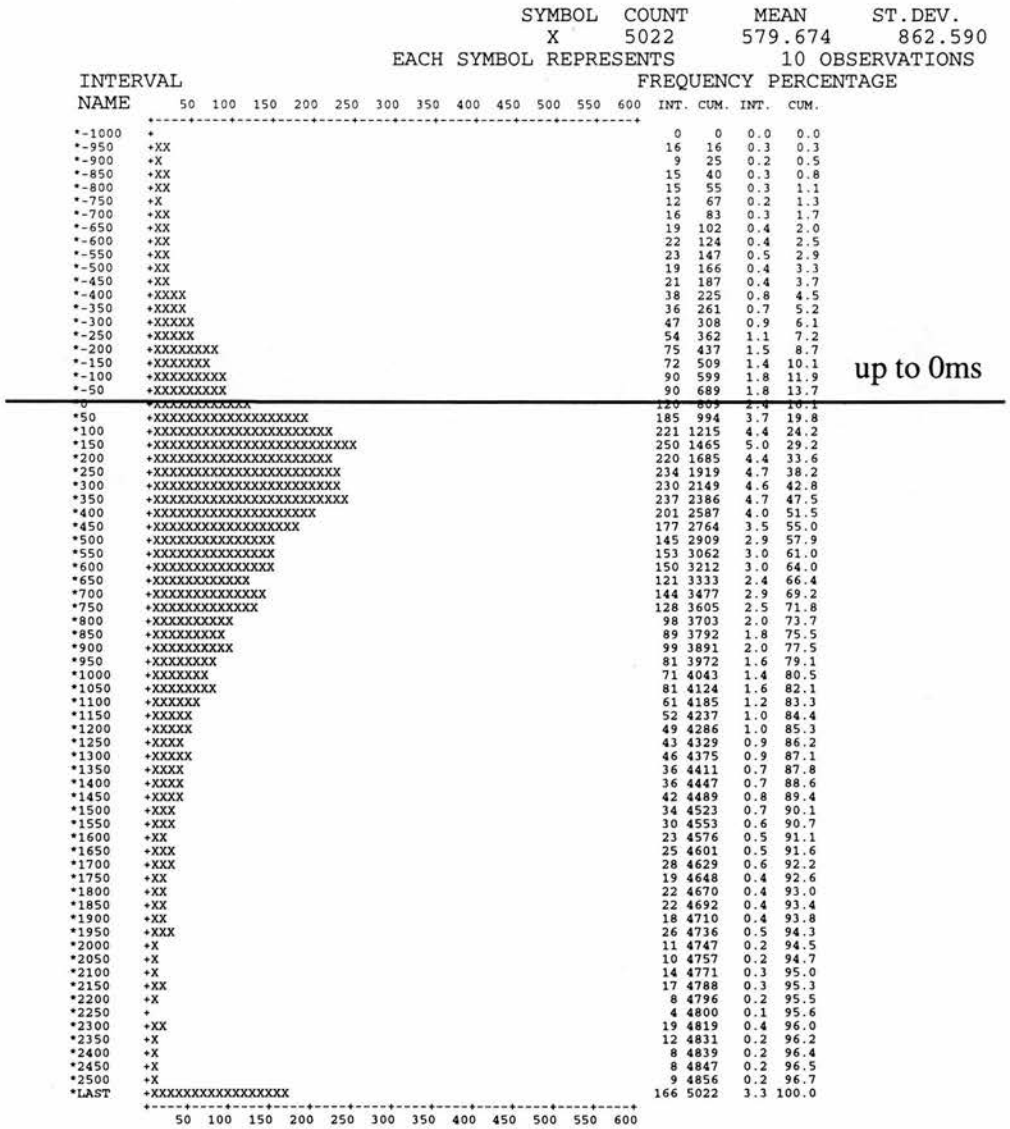
6.4.1.2 Eyecontact (2 levels)

The 5-way ANOVA crossing the independent variables Game Boundary, Eyecontact, Role, Familiarity, and Sex showed that the Eyecontact variable was significant (F (1, 10969) = 23.69, $p < 0.0001$). The hypothesis that the potential to be able to look at another speaker is significant in the timing of turn-taking can be accepted.

The mean inter-speaker interval for +Eyecontact dialogues (mean=579.7ms, SD=669.5ms, n=5995) is higher than for -Eyecontact dialogues (mean = 422.2ms, sd = 862.6ms, n = 5022). This shows that the potential to see the other participant does make a difference to mean inter-speaker interval durations. This may be because being able to see the other participant permits a greater tolerance of longer inter-speaker intervals than when there is no possibility of seeing the other speaker. Temporal coordination need not be as tight when participants are able to see each other. When it is not possible to see a partner in a conversation, mean inter-speaker interval durations are lower because interlocutors generally over-compensate, and become less tolerant of such longer inter-speaker intervals. The general 'window' of acceptable inter-speaker interval durations becomes smaller.

Figures 4 and 5 below show the distributions of exchanges separated into eyecontact and no-eyecontact. The distribution of the mean inter-speaker intervals of eyecontact exchanges is broader than no eyecontact exchanges (the standard distribution is greater), but that the proportion of negative exchanges is lower in eye contact exchanges (16.1% overlap) than in no eyecontact exchanges (18.9% overlap). There is also a relatively broad distribution of intervals in the former about the range 50-200ms.

Anderson et al. (1997) found that even when two speakers were able to see one another, much of the time both were looking down. They also found similar patterns in non-Map Task face-to-face interactions. In this respect, one might assume that there would be little or no significant difference between mean inter-speaker intervals in eye contact and no eye contact dialogues. But there is also the liklihood that the *potential* to see a partner may make a difference to the social context of the task.



SYMBOL COUNT MEAN ST.DEV.
X 5022 579.674 862.590
EACH SYMBOL REPRESENTS 10 OBSERVATIONS

INTERVAL NAME	50 100 150 200 250 300 350 400 450 500 550 600	FREQUENCY INT.	CUM.	PERCENTAGE INT.	CUM.
*-1000	+	0	0	0.0	0.0
*-950	+XX	16	16	0.3	0.3
*-900	+X	9	25	0.2	0.5
*-850	+XX	15	40	0.3	0.8
*-800	+XX	15	55	0.3	1.1
*-750	+X	12	67	0.2	1.3
*-700	+XX	16	83	0.3	1.7
*-650	+XX	19	102	0.4	2.0
*-600	+XX	22	124	0.4	2.5
*-550	+XX	23	147	0.5	2.9
*-500	+XX	19	166	0.4	3.3
*-450	+XX	21	187	0.4	3.7
*-400	+XXXX	38	225	0.8	4.5
*-350	+XXXX	36	261	0.7	5.2
*-300	+XXXXX	47	308	0.9	6.1
*-250	+XXXXX	54	362	1.1	7.2
*-200	+XXXXXXXX	75	437	1.5	8.7
*-150	+XXXXXXX	72	509	1.4	10.1
*-100	+XXXXXXXXX	90	599	1.8	11.9
*-50	+XXXXXXXXX	90	689	1.8	13.7
0	+XXXXXXXXXXXX	120	809	2.4	16.1
*50	+XXXXXXXXXXXXXXXXXXX	185	994	3.7	19.8
*100	+XXXXXXXXXXXXXXXXXXXXXX	221	1215	4.4	24.2
*150	+XXXXXXXXXXXXXXXXXXXXXXXXXX	250	1465	5.0	29.2
*200	+XXXXXXXXXXXXXXXXXXXXX	220	1685	4.4	33.6
*250	+XXXXXXXXXXXXXXXXXXXXXX	234	1919	4.7	38.2
*300	+XXXXXXXXXXXXXXXXXXXXXX	230	2149	4.6	42.8
*350	+XXXXXXXXXXXXXXXXXXXXXXXX	237	2386	4.7	47.5
*400	+XXXXXXXXXXXXXXXXXX	201	2587	4.0	51.5
*450	+XXXXXXXXXXXXXXXXX	177	2764	3.5	55.0
*500	+XXXXXXXXXXXXXX	145	2909	2.9	57.9
*550	+XXXXXXXXXXXXXX	153	3062	3.0	61.0
*600	+XXXXXXXXXXXXXX	150	3212	3.0	64.0
*650	+XXXXXXXXXXX	121	3333	2.4	66.4
*700	+XXXXXXXXXXXXX	144	3477	2.9	69.2
*750	+XXXXXXXXXXXX	128	3605	2.5	71.8
*800	+XXXXXXXXX	98	3703	2.0	73.7
*850	+XXXXXXXX	89	3792	1.8	75.5
*900	+XXXXXXXXX	99	3891	2.0	77.5
*950	+XXXXXXX	81	3972	1.6	79.1
*1000	+XXXXXX	71	4043	1.4	80.5
*1050	+XXXXXXX	81	4124	1.6	82.1
*1100	+XXXXX	61	4185	1.2	83.3
*1150	+XXXX	52	4237	1.0	84.4
*1200	+XXXX	49	4286	1.0	85.3
*1250	+XXX	43	4329	0.9	86.2
*1300	+XXXX	46	4375	0.9	87.1
*1350	+XXX	36	4411	0.7	87.8
*1400	+XXX	36	4447	0.7	88.6
*1450	+XXX	42	4489	0.8	89.4
*1500	+XXX	34	4523	0.7	90.1
*1550	+XXX	30	4553	0.6	90.7
*1600	+XX	23	4576	0.5	91.1
*1650	+XXX	25	4601	0.5	91.6
*1700	+XXX	28	4629	0.6	92.2
*1750	+XX	19	4648	0.4	92.6
*1800	+XX	22	4670	0.4	93.0
*1850	+XX	22	4692	0.4	93.4
*1900	+XX	18	4710	0.4	93.8
*1950	+XXX	26	4736	0.5	94.3
*2000	+X	11	4747	0.2	94.5
*2050	+X	10	4757	0.2	94.7
*2100	+X	14	4771	0.3	95.0
*2150	+XX	17	4788	0.3	95.3
*2200	+X	8	4796	0.2	95.5
*2250	+	4	4800	0.1	95.6
*2300	+XX	19	4819	0.4	96.0
*2350	+X	12	4831	0.2	96.2
*2400	+X	8	4839	0.2	96.4
*2450	+X	8	4847	0.2	96.5
*2500	+X	9	4856	0.2	96.7
*LAST	+XXXXXXXXXXXXXXXX	166	5022	3.3	100.0

up to 0ms

Figure 4 - Distribution of inter-speaker intervals for eye contact exchanges

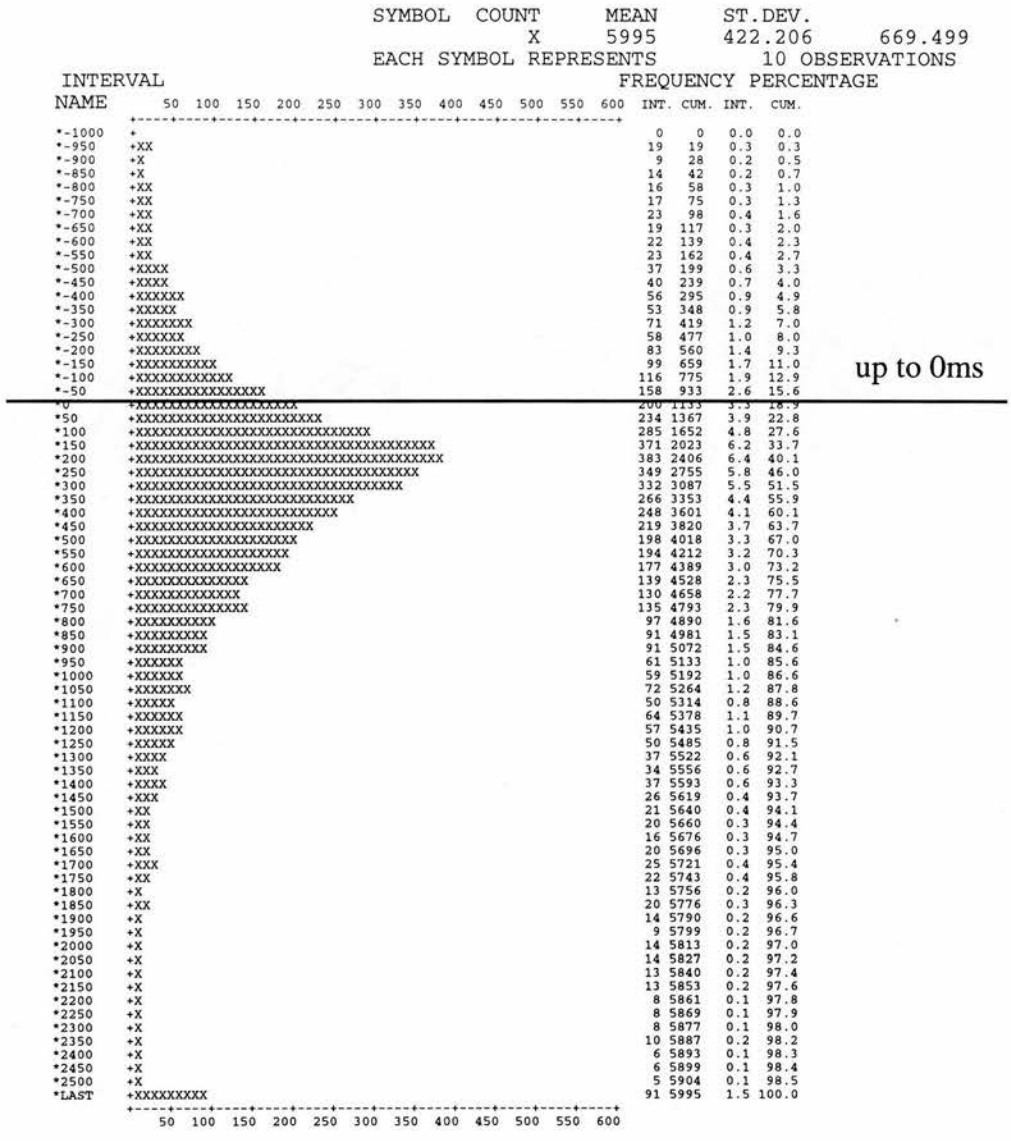


Figure 5 - Distribution of inter-speaker intervals for no eye contact exchanges

6.4.1.3 Conversational Game Boundary (2 levels)

The 5-way ANOVA crossing the independent variables Game Boundary, Eyecontact, Role, Familiarity, and Sex showed that the Game Boundary variable was significant (F (1, 10969) = 131.31, $p < 0.0001$). There are differences in inter-speaker intervals according to whether there was a game boundary in an exchange or not.

The presence of a game boundary in an exchange is associated with a significantly longer inter-speaker interval (n = 3090, mean = 656ms, sd = 843ms) than is associated with non-game boundary speaker switches (n = 7927, mean = 431ms, sd = 726ms).

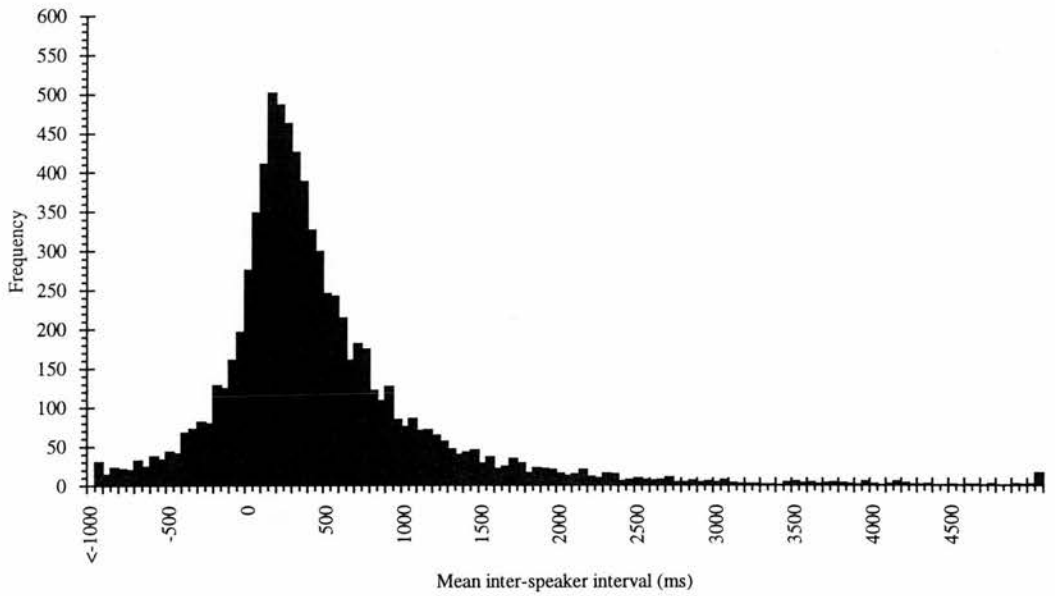


Figure 6 - The distribution of inter-speaker intervals game-internally. n = 7927, mean = 431ms, sd = 726ms (with data split into 50ms bins)

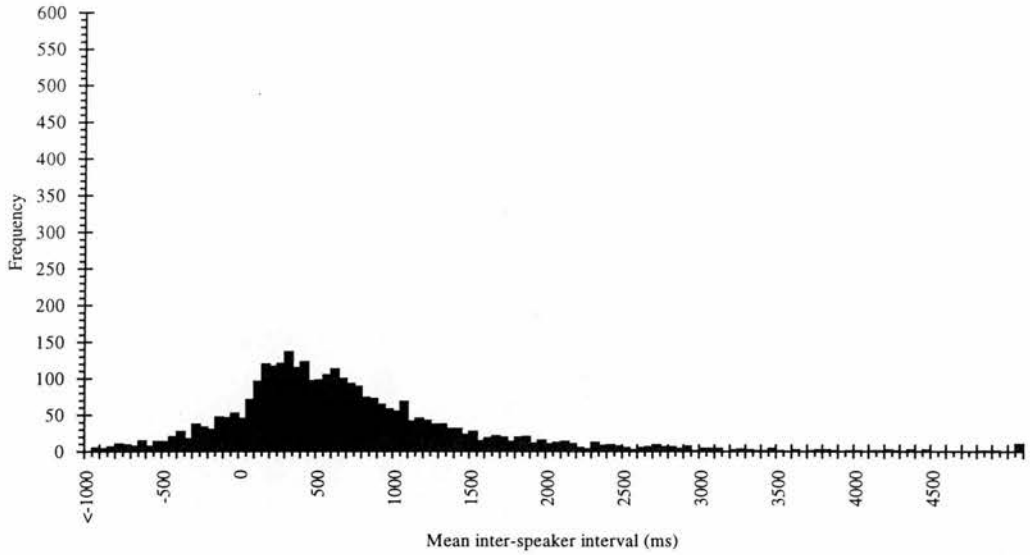


Figure 7 - The distribution of inter-speaker intervals at game boundaries. n = 3090, mean = 656ms, sd = 843ms (with data split into 50ms bins)

Inter-speaker intervals at game boundaries have a greater mean and standard deviation than intervals in game-internal exchanges. Game boundary exchanges therefore have a broader distribution of inter-speaker intervals, but a lower proportion of overlapped exchanges (game-internal = 18.9% overlap; game boundary = 14.3% overlap). It would appear that intervals at game boundaries are subject to less rigid temporal restrictions than game-internal intervals.

One reason for the difference between inter-speaker intervals at game boundaries and game-internal inter-speaker intervals might be that new games are more often than not started by a giver. In other words, the difference may be a disguised effect of the role of the speakers. But as Figure 8 below shows, this is not the case. Mean intervals are consistently greater for giver-follower exchanges than for follower-giver exchanges, irrespective of whether the exchange occurs across a boundary, or within a game.

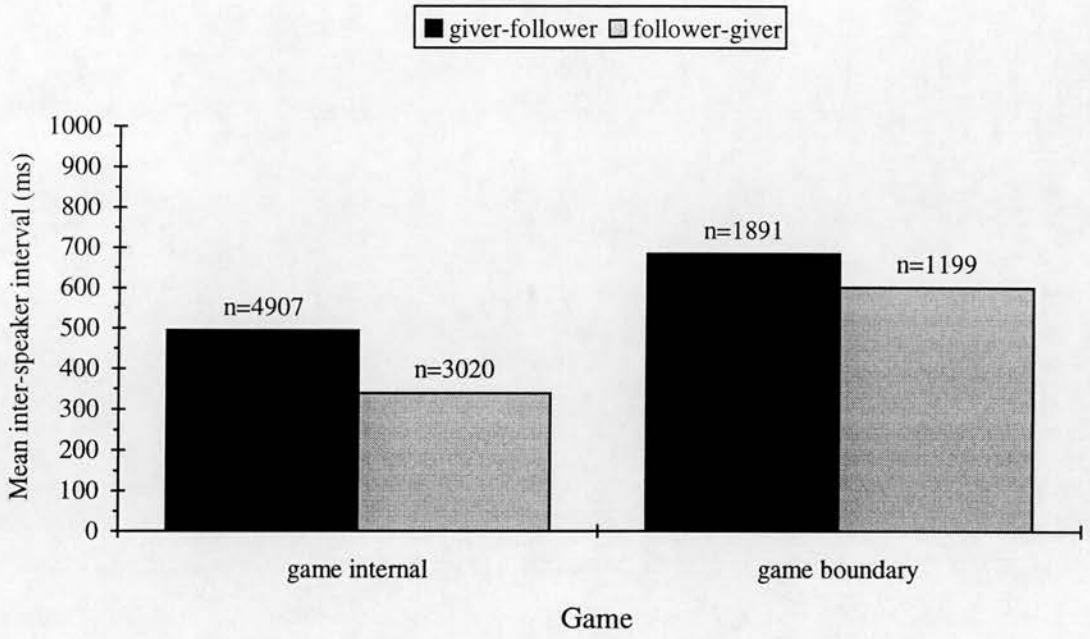


Figure 8 - Mean inter-speaker interval durations by game category

There are therefore genuine effects associated with a distinction between a game boundary and a non-game boundary, which reflect differences in processing time when a new part of the conversation is started.

6.4.1.4 Familiarity (2 levels) and Sex (3 levels)

The 5-way ANOVA crossing the independent variables Game Boundary, Eyecontact, Role, Familiarity, and Sex showed that the main effects for Familiarity and for Sex were not significant. The following interactions were significant however: a) Familiarity x Sex (F (2, 10969) = 6.74, p =0.0012); b) Familiarity x Sex x Eyecontact (F (2, 10969) = 18.65, $p < 0.0001$); c) Familiarity x Sex x Eyecontact x Game boundary (F (2, 10969) = 4.36, p = 0.0128). There is an interaction effect between the Familiarity and Sex variables, which is possibly a result of the imbalance between the gender of participants, and their relationship with their partners (as noted in section 6.2.2).

6.4.2 Test of the Significance of the Match, Route, and Contrast Variables with Respect to Inter-Speaker Interval Using a 6-way ANOVA of Game Boundary x Eyecontact x Role x Match x Route x Contrast

Game Boundary - 2 levels. Boundary, non-boundary.

Eyecontact - 2 levels. Potential for eyecontact, no potential for eye contact.

Role - 2 levels. Giver-follower, follower-giver.

Contrast - 2 levels. Contrast in main features on giver's map, and no contrast.

Match - 2 levels. Matching contrast, no matching contrast.

Route - 4 levels. Four different types of route.

The map design was reflected in three variables: Match, Contrast, and Route Number. The Game boundary, Eyecontact, and Role main effects were all significant, as the ANOVAs in section 4.1 showed. This 6-way ANOVA also showed that two of the variables associated with map design were significant - Match (F (1, 10889) = 14.66, $p < 0.0001$), and Route (F (3, 10889) = 3.00, $p < 0.0292$).[26]

However, interactions between Route and Eyecontact (F (3, 10889) = 4.99, p = 0.0018), Route and Match (F (3, 10889) = 6.24, p = 0.0003), and Route and Contrast (F (3, 10889) = 2.70, p = 0.0439), indicated that the significance of the Route variable may result from other variables - as supported by a further (1-way) ANOVA with Route as the independent variable. Route was found not to be significant with respect to inter-speaker interval duration. Route was ignored in the rest of the analysis.

A further analysis of the Match variable showed that the mean inter-speaker interval for +Match exchanges was 469ms (n = 5816, sd = 718ms), whereas the mean interval associated with -Match exchanges was 522ms (n = 5201, sd = 819ms). Therefore, when there was a match in the arrangement of contrasting pairs of landmarks of giver and follower's maps, the mean inter-speaker interval was significantly shorter than when there was no match.

---

[26]See Appendix C for a full listing of the significant results in this ANOVA.

These results explain the observed differences between mean inter-speaker intervals of different conversation numbers, bearing in mind that the odd-numbered conversations were all +Match, and the even conversations were all -Match. Each of the odd-numbered conversations had a lower mean inter-speaker interval than its following even-numbered conversation. That is, conversation 1 had a lower mean interval than conversation 2. Conversation 3 had a lower mean interval than conversation 4, and so on, as shown in Figure 9 below.

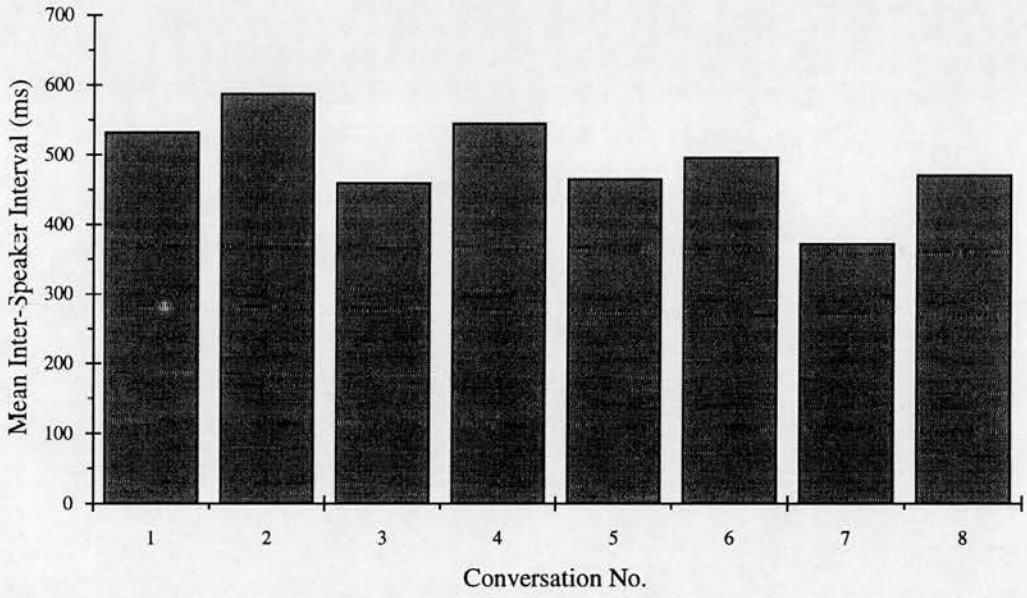


Figure 9 - Mean inter-speaker intervals for each conversation number

Differences between maps result from differences in the Contrast, Match, and Route variables. Generally the significance of the map variable is caused by the Match variable. Contrast and Route are not significant. When the contrast between the master features on the giver's and follower's maps does not match (for example, when the giver has diamond mine and gold mine, but the follower has two gold mines), there should be greater difficulty with the task than when there is a match. It would require more planning time in a -Match case than in a +Match case, and the mean inter-speaker intervals should be greater. In effect, this result shows that even small changes in the underlying task that participants are involved in (here, trying to solve a problem) will be reflected in inter-speaker intervals.

6.4.3 Test of the Significance of the Task Familiarity Variable with Respect to Inter-Speaker Interval Using a 5-way ANOVA - Game Boundary x Eyecontact x Role x Match x Task Familiarity

Game Boundary - 2 levels. Boundary, non-boundary.

Eyecontact - 2 levels. Potential for eyecontact, no potential for eye contact.

Role - 2 levels. Giver-follower, follower-giver.

Match - 2 levels. Matching contrast, no matching contrast.

Task Familiarity - 2 levels. Conversation numbers 1 and 2, conversation nos. 3 to 8.

This 5-way ANOVA showed that Task Familiarity played a significant role in the determination of inter-speaker interval duration (F (1, 10985) = 26.18, $p < 0.0001$).[27] In other words, the mean inter-speaker intervals of conversations 1 and 2 (where participants had not before taken part in a Map Task dialogue) were significantly different from the mean intervals of conversations 3 - 8 (where participants had taken part in at least one dialogue before). The mean inter-speaker intervals for conversations 1 and 2 in each quad were longer (mean=561ms, sd = 871ms, n = 3140) than for conversations 3-8 (mean = 467ms, SD=721, n = 7877). The greater standard deviation of conversations 1 and 2 suggests that on first presentation of the task, participants had more variable inter-speaker intervals.

Figure 9 in section 1.4.2 above shows a slight decline in mean inter-speaker interval from conversation number 1 through to conversation number 8. Again, this shows that participants tend to become more efficient and effective the more familiar they become with a task, and as a result the inter-speaker intervals become shorter. Figure 9 also clearly shows a marked 'stepping' in the series of mean inter-speaker intervals for the conversations. This is caused by the effect of the Match variable - the odd-numbered conversations are +Match; the even-numbered conversations are -Match.

16 different maps were used in the Map Task. 4 different maps were used in each of quads 1-4, with the same maps used in quads 5-8. In each quad, each map

---

[27]See Appendix C for a listing of all the other significant results in this ANOVA.

was used twice. On its second use, the giver was familiar with the map. If familiarity with the map and with the task required to guide someone else through it makes for greater efficiency, then mean inter-speaker interval duration should be lower in the last four dialogues of any quad than in the first four (bearing in mind that in the last four dialogues, the giver was familiar with the map). The mean inter-speaker interval of conversations 1 - 4 was 532ms (n = 5968, sd = 797ms), whereas the mean inter-speaker interval of conversations 5 - 8 was 450ms (n = 5049, sd = 729ms). A one-way ANOVA (with this 2-way split of conversation number as the independent variable) showed that there was a significant difference in the mean inter-speaker intervals of conversations 1-4 and conversations 5-8 (F (1, 11015) = 31.34, $p <$ 0.0001).

There was therefore good evidence that familiarity with a task is reflected in reduced inter-speaker interval duration, caused by the need for less planning and decision time by both participants.

6.4.4 4-way ANOVA - 2-class a-move[28] x 2-class b-move x Eyecontact x Role.

2-class a-move - 3 levels. Initiating a-moves, responding a-moves.
2-class a-move - 3 levels. Initiating b-moves, responding b-moves.
Eyecontact - 2 levels. Potential for eyecontact, no potential for eye contact.
Role - 2 levels. Giver-follower, follower-giver.

The 2-class categorization of moves is similar to the categorization of exchanges into game-internal and game-boundary exchanges. The latter distinction has already been shown to be significant in section 6.4.1.3. As noted earlier, moves were originally classified as either initiators, responses, or transitions. Transition moves were not common, particularly as a-moves (for transition/initiator combinations, n=36. For transition/response combinations, n=20. And for transition/transition combinations, n=6). In this analysis they were treated as initiators rather than being eliminated

---

[28]Some moves were also classified as *transitional* moves. This category was omitted because of insufficient data to allow the ANOVA to be calculated.

entirely. Transitional moves (a class made up entirely of *ready* moves) were more likely to initiate a new game.

I could not cross the 2-class move variable with all five variables already shown to be significant: Game Boundary, Eyecontact, Role, Match, and Task Familiarity. This produced many empty cells. I carried out an ANOVA which crossed a-move, b-move, Eyecontact, and Role. This showed that the 2-class b-move variable was significant (F (1, 10184) = 131.38, $p < 0.0001$). It is therefore important whether the second utterance in an exchange starts with an initiator or a response-type move.

Exchanges had longer mean inter-speaker intervals where b-moves were initiators than where b-moves were responses. This is shown in Figure 10 below, where the black bars represent exchanges with initiator b-moves. Grey bars represent response b-moves.

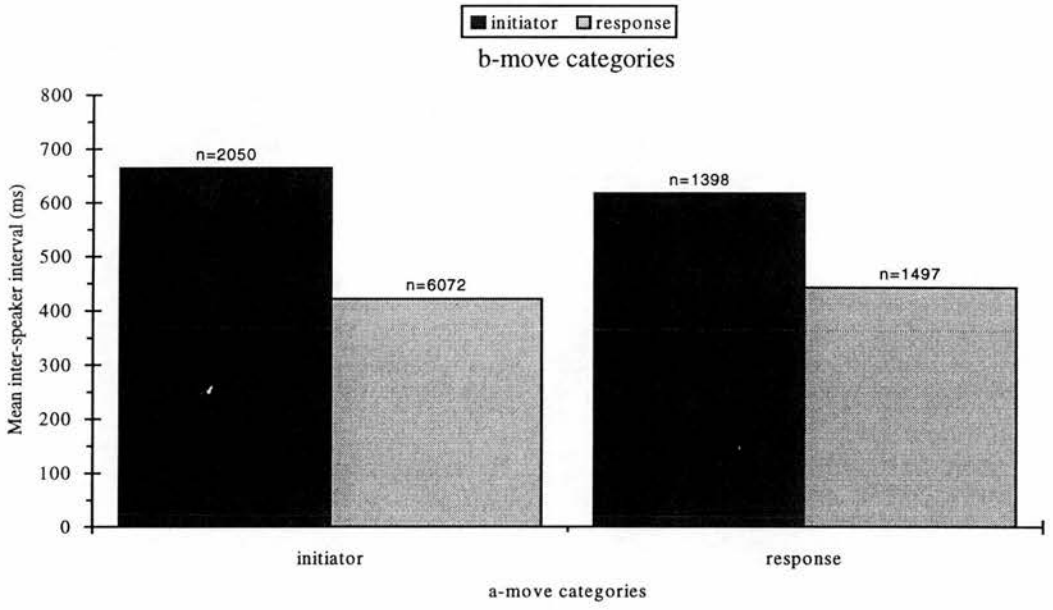


Figure 10 - The mean inter-speaker intervals associated with different categories of the 3-class move variable, for a-move and b-move.

These results agree with the game-boundary results. Since (generally speaking) initiator moves have the potential to start new games, and since initiator b-moves are significantly longer than response b-moves, we have further evidence that at least the potential to start a new game can result in a longer inter-speaker interval. Generally, the second move in an exchange has more significance to inter-speaker interval duration than does the first move. A highly significant factor in inter-speaker interval

duration is therefore whether a new game has been started or not. The reasons for this, assuming the principle of processing time (Clark, 1996), are that to start a new game on average requires more thought and planning on the part of the speaker than a game-internal exchange.

6.4.5 Test for the Significance of the 12-class-a-move Variable with Respect to Inter-Speaker Interval Using a 2-way ANOVA - 6-class-a-move x 9-class-b-move

6-class a-move - 6 levels. *check, explain, instruct, reply-y, reply-w, query-yn*
9-class b-move - 9 levels. *acknowledge, align, check, explain, instruct, ready, reply-y, query-yn, query-w*

Note that *acknowledge* moves were not used in this analysis in the a-move position because of the constraints on backchannelling and responses to backchannel signals (as mentioned in Chapter 4).

Even the reduced sets of moves would allow only a relatively limited analysis because of empty cells (see Appendix B2). An ANOVA crossing six a-moves with nine b-moves showed that both a-move (F (5, 7304) = 4.21, p = 0.0008) and b-move (F (8, 7304) = 4.70, $p < 0.0001$) were significant in explaining differences in inter-speaker intervals. But there was no significant interaction. Differences in inter-speaker intervals can therefore be partly explained by move category, and partly by the category of move which follows a speaker switch.

A Tukey test indicated that (at the 0.01 level of significance) there were 5 main move-pairs which had mean inter-speaker intervals which were different from the mean intervals of other pairs. These were:

check/reply-y	mean = 176ms, n = 868
explain/acknowledge	mean = 366ms, n = 623
query-yn/reply-y	mean = 328ms, n = 640
instruct/ready	mean = 939ms, n = 98
reply-w/query-w	mean = 1166ms, n = 26

Figures 11 and 12 below show by way of example those move-pairs which differ significantly from the *check/reply-y* pair and the *instruct/ready* pair.

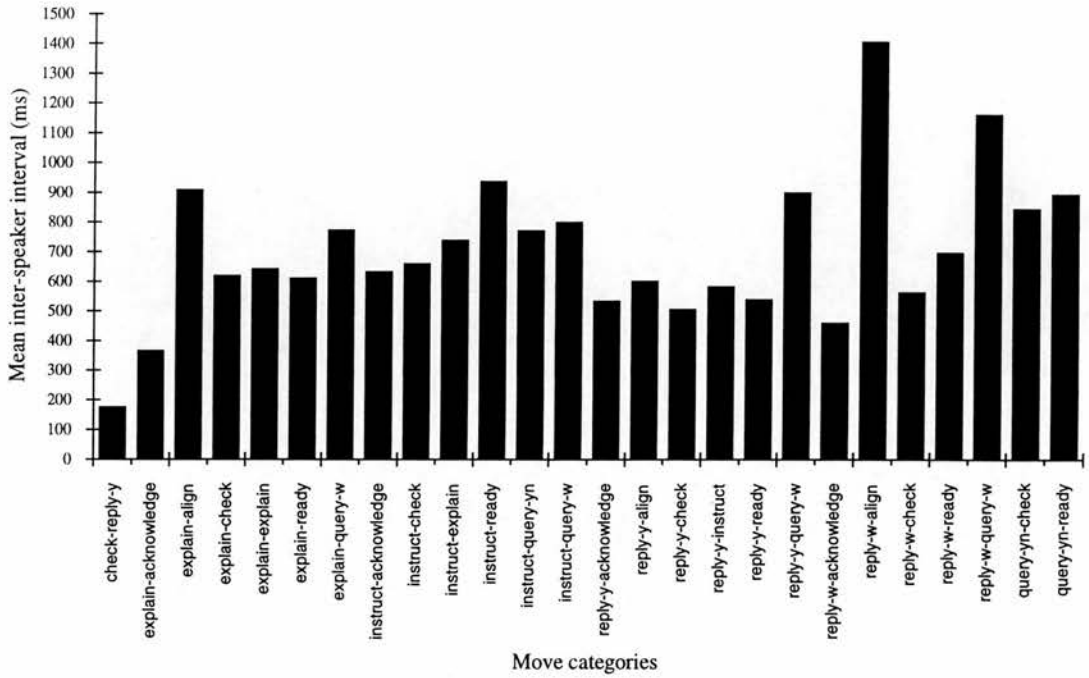


Figure 11 - Mean inter-speaker intervals of those move pairs which differ significantly from the *check/reply-y* pair (which is shown on the far left)

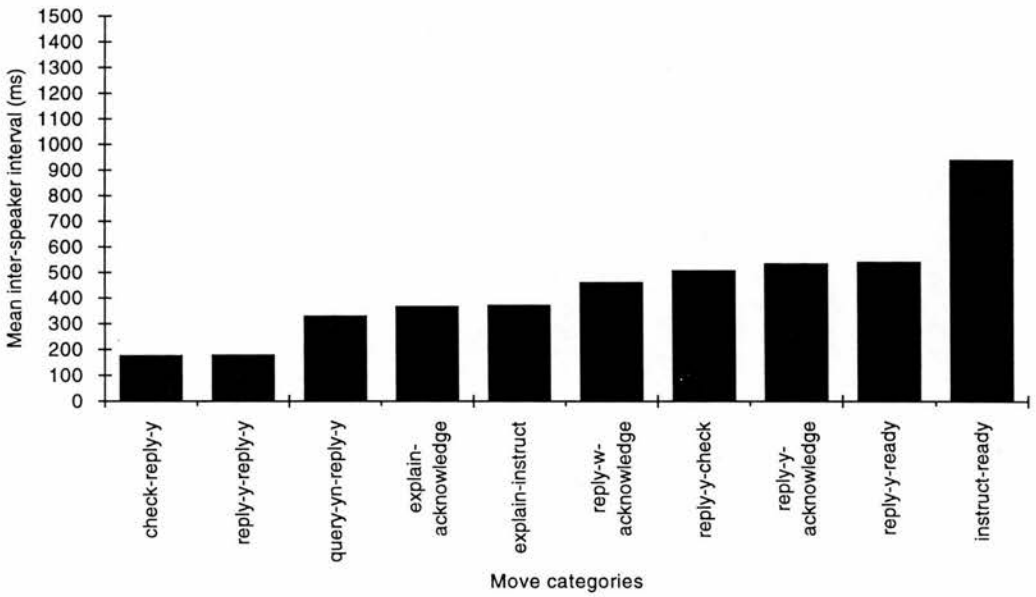


Figure 12 - Mean inter-speaker intervals of those move pairs which are significantly different from the *instruct/ready* move pair (shown at the far right of the histogram)

There are three points to be made about these results. First, the three move pairs with low means (*check/reply-y*, *query-yn/reply-y*, and *explain/acknowledge*) also have relatively high cell sizes (868, 640, and 623 respectively). If these pairs are relatively common, one might therefore suppose that they represent unmarked move-pairs. For example, it is more 'natural' for a *reply-y* move to follow a *check* move, or for a *reply-y* move to follow a *query-yn* move. In other words, unmarked cases of move-pairs may also have relatively short mean inter-speaker intervals.

Second, the three short move pairs also occur in game-internal position, and the second move in each pair is not a game-initiating move. It therefore appears that, to some extent at least, these results reflect game boundary effects.

Third, it is possible that the high mean of *instruct/ready* move pairs (939ms) may result from drawing time. That is, such move pairs are characterised by an instruction from the Information Giver, and then (on average almost a second later) an utterance such as '*OK*' or '*Right*' by the Information Follower. Importantly, *ready* moves differ from many *acknowledge* moves in that *ready* moves tend to end a game, or act as a transition between games (see Chapter 2). They are used, at least within the goal-oriented confines of the Map Task, to say "OK, I've understood and finished the task you gave me to do". An *acknowledge* move, however, typically is used to signal that an utterance was understood. Because of the uses that a speaker will make of a *ready* move, it is likely that many will follow some activity that results from an instruction in the Map Task - for example drawing a part of a route on the map. This returns us to the problem mentioned in Chapter 4, section 7, where many positive inter-speaker intervals result from physical limitations, such as how long it takes to draw a line of a certain length.

This analysis was not exhaustive. The small sample sizes for some categories of move-pairs meant that a full ANOVA could not be carried out, and that certain move categories had to be eliminated altogether from the main analysis. This meant that some common combinations of move (for example, *align/reply-y*, n = 730) were left out of the analysis because one of the moves in the pair had been eliminated. And because they were left out of the analysis it was not possible to find out if they had means which were significantly larger or smaller than other move categories.

However, it was possible to 'retrieve' some move pairs, and to determine if they had particularly long or short mean inter-speaker intervals. If a pair's mean inter-speaker interval was below 400ms, it was classed as a 'short' pair. If a pair's mean inter-speaker interval was above 800ms, it was classed as a 'long' pair. These cut-off points were based on the mean intervals of move pairs involved in the significant differences listed above. The list of pairs with particularly 'long' and 'short' mean inter-speaker intervals are shown in Table 2 (where the 'retrieved' move pairs are indicated with an asterisk).

Table 2 - Mean intervals, standard deviations, and numbers for move pairs

Move Pairs	**'short'**			**'long'**		
	mean (ms)	SD (ms)	n	mean (ms)	SD (ms)	n
check/reply-y	176	437	868			
**reply-y/reply-y*	178	342	21			
**align/reply-y*	254	469	730			
**check/clarify*	261	495	152			
query-yn/reply-y	328	524	640			
**query-yn*/ackn	330	697	21			
**check/reply-w*	338	548	46			
**clarify/acknowledge*	356	669	427			
explain/acknowl'ge	366	714	623			
**explain/instruct*	371	752	45			
**explain/reply-y*	389	764	22			
**reply-n*/acknowl'ge	394	494	180			
**instruct/query-w*				800	1013	144
**query-yn/check*				847	713	28
**reply-y/query-w*				901	1257	47
**explain/align*				909	1582	24
instruct/ready				939	1105	98
reply-w/query-w				1166	1830	26

We can see, for example, that the class of *align/reply-y* moves is common, and has a low mean inter-speaker interval. Although no definite proof exists of whether this class is significantly smaller than many other classes, Table 1 helps indicate that it might be.

A one-way ANOVA of *all* the original a-move categories (the *acknowledge* move had been eliminated because of the criteria set out in chapter 4), and not just the 6 categories used above, revealed significant differences among the categories (F (10, 11006) = 34.64, $p < 0.0001$). A Tukey test of a-moves revealed several main significant differences between move categories. Figure 13 shows those a-moves which were found to be significantly different from *align* and *check* a-moves (the latter shown as black bars). Figure 14 shows those a-moves which were found to be significantly different from *instruct* and *ready* a-moves (the latter are shown as black bars).

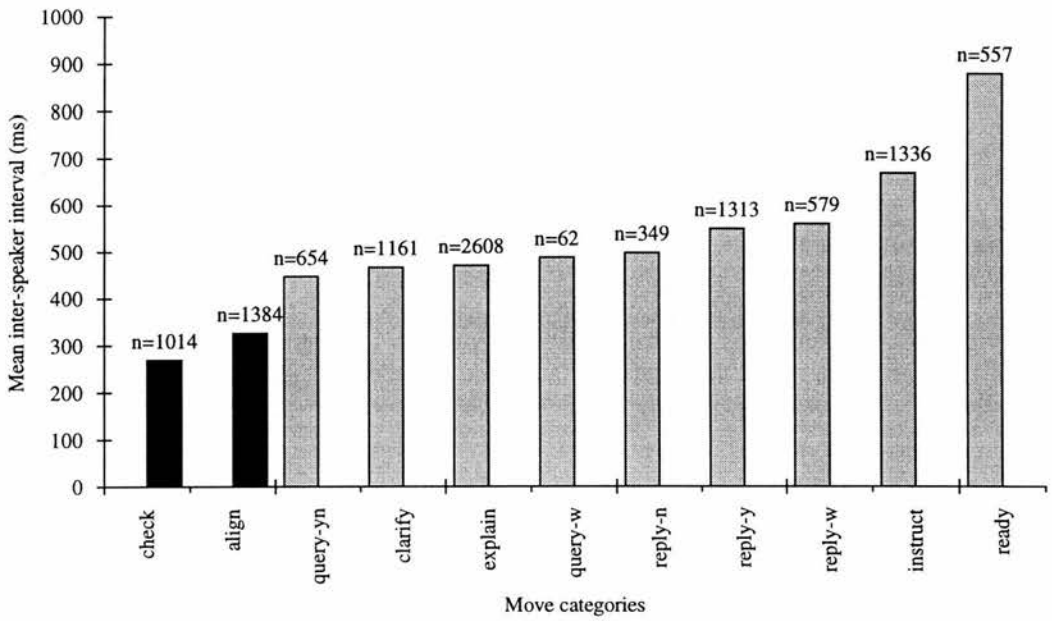


Figure 13 - Mean inter-speaker intervals of a-move categories which are significantly different from *align* and *check* a-moves

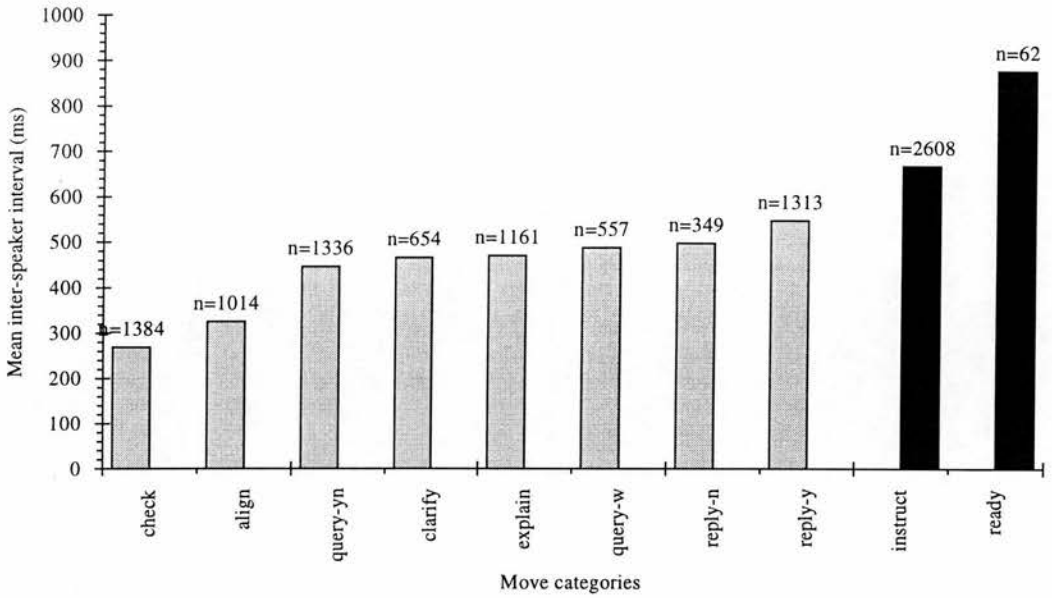


Figure 14 - Mean inter-speaker intervals of a-move categories which are significantly different from *instruct* and *ready* a-moves

These results show that *align* and *check* a-moves are generally followed by relatively short inter-speaker intervals, whereas *instruct* and *ready* a-moves are generally followed by relatively long inter-speaker intervals. It should be pointed out, however, that there are only 62 *ready* a-moves, so that the statistics for this category may not be totally reliable.

A similar one-way ANOVA of all the 12 original categories of the b-move variable, rather than the 9 used above, revealed a significant difference among the levels (F (11, 11005) = 38.41, $p < 0.0001$). I carried out a Tukey test on the mean inter-speaker intervals preceding b-move categories. The pattern of significant differences was slightly more complex than for a-moves. A Scheffé test was therefore carried out to select only the more important differences. Figures 15, 16, and 17 summarise the main differences.

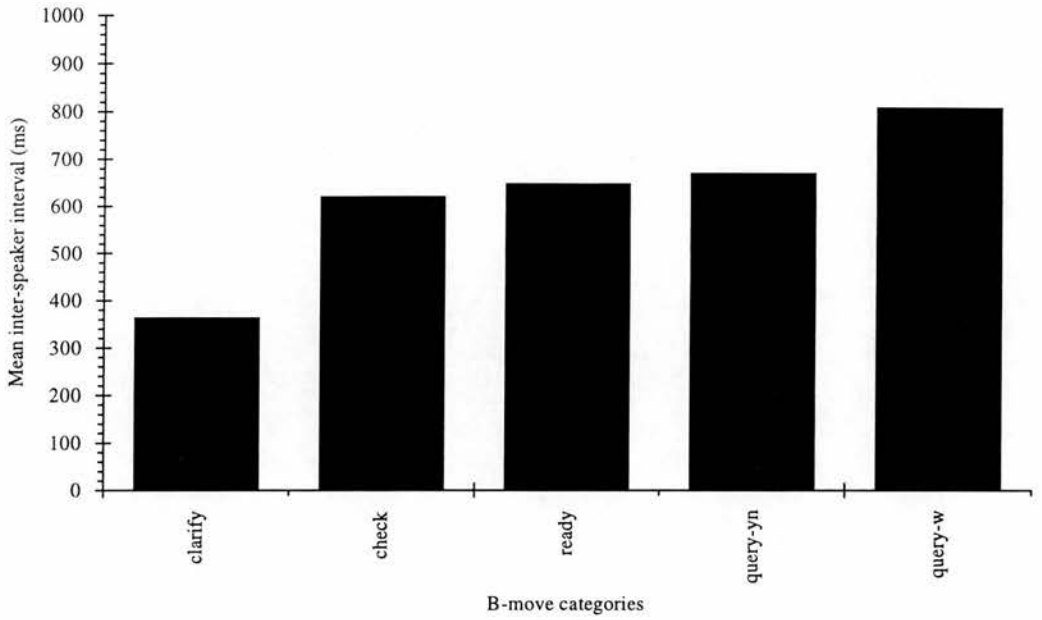


Figure 15 - Mean inter-speaker intervals preceding those b-move categories which have significantly greater intervals than the intervals preceding *clarify* moves.

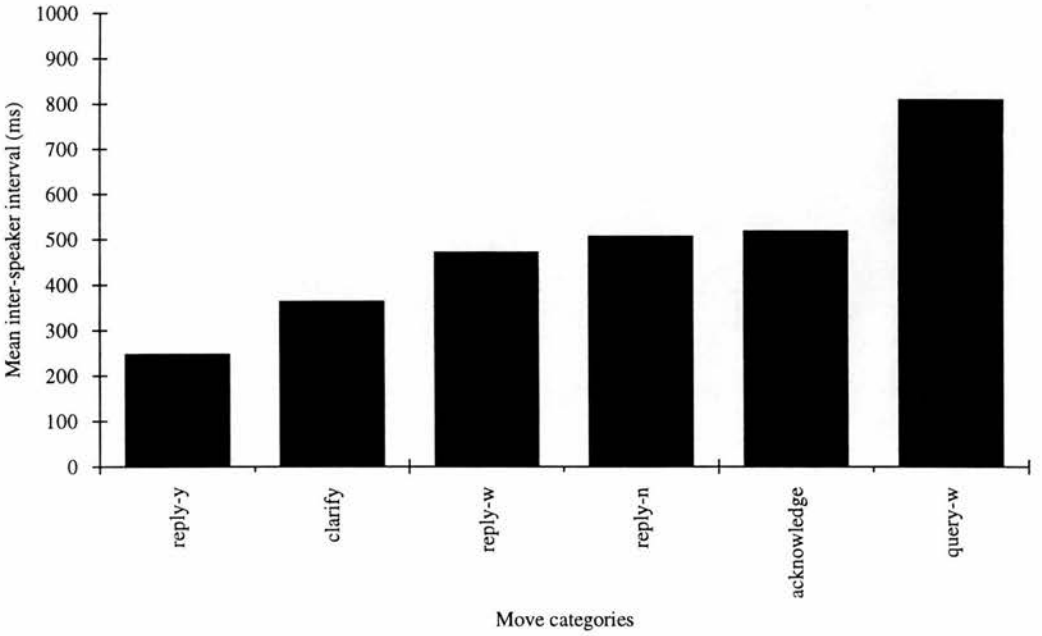


Figure 16 - Mean inter-speaker intervals preceding those b-move categories which have significantly smaller intervals than the intervals preceding *query-w* moves.

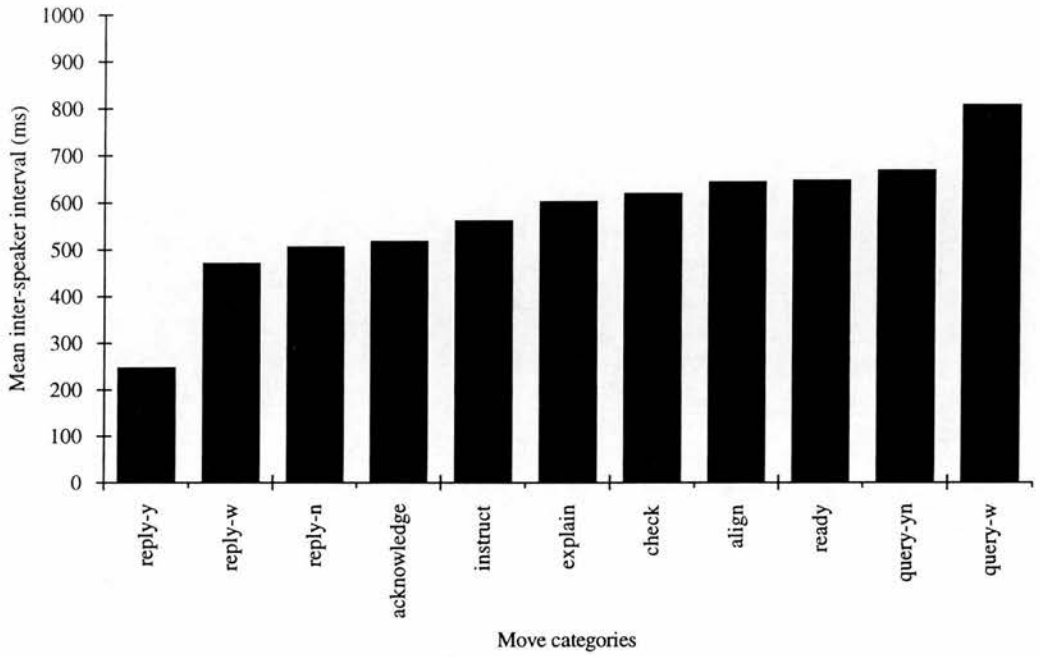


Figure 17 - Mean inter-speaker intervals preceding those b-move categories which have significantly greater intervals than the intervals preceding *reply-y* moves.

If the inter-speaker intervals which precede b-moves are classified into two broad classes - 'long' and 'short' - then it can be seen that generally *clarify* and *reply-yn* b-moves are preceded by relatively short inter-speaker intervals. *Query-w* b-moves are generally preceded by relatively long inter-speaker intervals.

There are several points which become apparent:

1) The inter-speaker interval preceding a *reply-y* move tends to be relatively short. Of the 12 'short' move pairs, 5 have *reply-y* b-moves. Moreover, the short intervals preceding *reply-y* moves occur most frequently in *align/reply-y*, *check/reply-y*, and *query-yn/reply-y* move pairs.

4 of the 12 'short' move pairs have *acknowledge* b-moves. *Acknowledge* b-moves therefore tend to be preceded by relatively short inter-speaker intervals. Short

intervals preceding *acknowledge* moves are most common in *clarify/acknowledge*, *explain/acknowledge*, and *reply-n/acknowledge* move pairs.

2) *Explain* and *check* a-moves tend to be followed by relatively short inter-speaker intervals. Particularly common short move pairs with these a-moves are *check/clarify*, *check/reply-y*, and *explain/acknowledge* pairs.

3) *Instruct* and *ready* a-moves tend to be followed by relatively long inter-speaker intervals. *Instruct/query-w* and *instruct/ready* move pairs are particularly common.

4) *Query-w* b-moves tend to be preceded by relatively long inter-speaker intervals. *Reply-y/query-w*, *reply-w/query-w*, and *instruct/query-w* pairs are common examples of this. The latter move pair may also tend to be accompanied by a relatively long inter-speaker interval because of the *instruct* a-move.

It is difficult to assess the exact relationship between a-moves and b-moves. In general, b-moves have greater significance with respect to inter-speaker interval duration than a-moves. It is likely therefore that inter-speaker intervals may be relatively long whenever the b-move is a *query-w* move. This is true even where the a-move is, for example, a *check* move, where *check* a-moves are typically followed by relatively short inter-speaker intervals (mean = 269ms). On the other hand, some b-moves, such as the *reply-y* move, tend to be preceded by relatively short inter-speaker intervals, irrespective of the a-move that precedes them.

6.4.6 A Test of the Significance of the Gaze Variable with Respect to Inter-Speaker Interval Using a 1-way ANOVA

Gaze - 8 levels. Speaker as giver, listener as follower: a) speaker and listener look down; b) speaker looks up/listener looks down; c) listener looks up/speaker looks down; d) speaker and listener look up. Another four levels with Speaker as follower, listener as giver.

Because of the restriction of sample size for the gaze variable (only four of the Eyecontact quads were available for analysis) a full analysis was not possible. Gaze was categorised into four levels - *both participants look down*, *only speaker looks up*, *only listener looks up*, and *both look up*. A 1-way ANOVA with Gaze as the independent variable yielded a significant result (F (7, 2448) = 5.2, $p < 0.0001$). However, closer inspection of the mean inter-speaker intervals for each of the gaze categories showed that there was a split between those mean intervals falling within giver-follower exchanges and those falling within follower-giver exchanges (as would be expected from the giver-follower results noted above). Yet there was little difference between the mean intervals between gaze categories (see Figure 18).

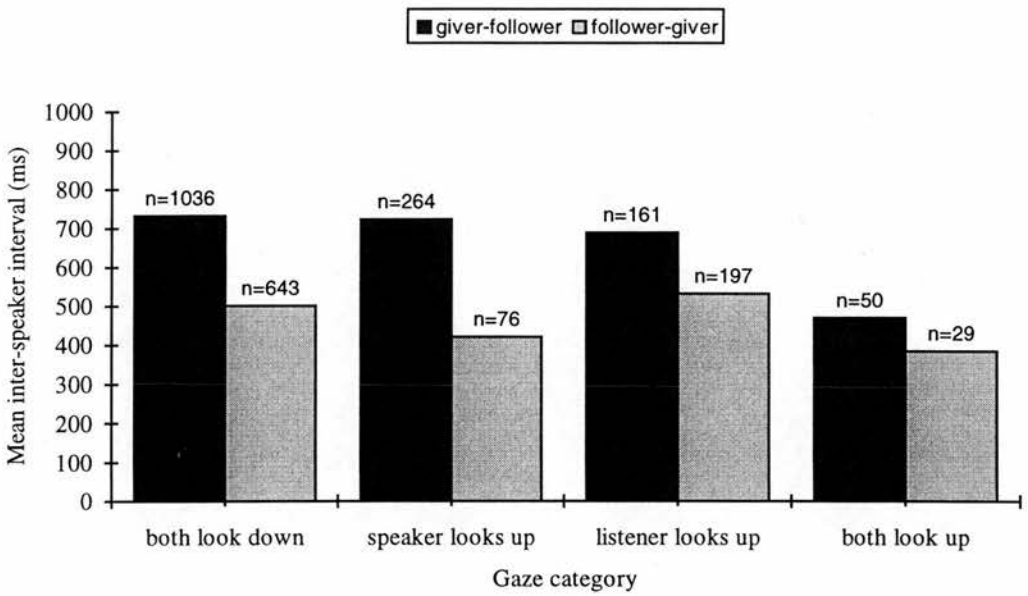


Figure 18 - Mean inter-speaker intervals by gaze category

I carried out a 2-way ANOVA, using Gaze and Role variables. This showed that only the Role variable was significant (F (1, 2448) = 7.96, p = 0.0048). A Tukey test showed only two significant differences. These were: a) between giver-follower and follower-giver exchanges where both participants were looking down; b) between giver-follower exchanges where the speaker was looking up, and follower-giver exchanges where both participants were looking down. Apparent Gaze effects are therefore caused by Role differences.

It should also be noted that of the 2456 exchanges that could be coded for gaze, 1679 (68.4%) involved no gaze at all. Only 79 exchanges (3.2%) involved mutual gaze. This is in accordance with (Anderson et al., 1997) which pointed out that even when there is the possibility of eye contact, speakers mostly look down, or away from one another. If gaze has any conversational function (which it may do) then it would not appear to be reflected in any direct manner in inter-speaker interval durations.

6.4.7 A Test of the Significance of the Shared Landmarks Variable with Respect to Inter-Speaker Interval Using a 1-way ANOVA

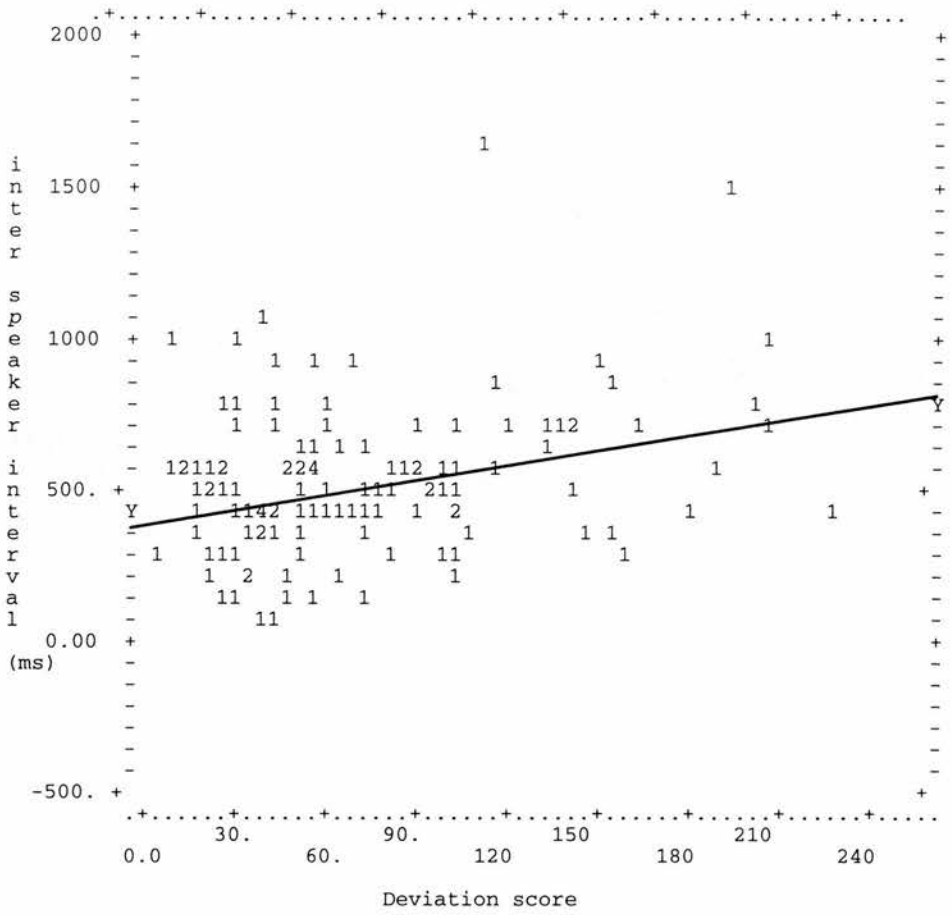
Shared landmarks - 2 levels. New and shared landmarked, new and unshared landmark.

The mean inter-speaker interval associated with exchanges where a new and shared landmark was introduced was 525ms (n = 74, sd = 653ms). The mean interval associated with exchanges where a new, unshared landmark was introduced was 648ms (n = 89, sd = 582ms). This, however, was not a significant difference - presumably because of the relatively small sample size.

6.4.8 Deviation Score

Deviation score was used as an approximate measure of the degree of difficulty encountered by participants in a dialogue. The greater the deviation score for a particular dialogue, the greater the assumed difficulty that the participants had in completing the task. It was predicted that, if the basic assumption of the principle of processing time was correct, there should be a correlation between deviation score and mean inter-speaker interval.

There was a small yet significant correlation between deviation score and mean inter-speaker interval (R = 0.279, p = 0.002). Figure 19 below shows this correlation.



R = .279
p =.002
mean = 72.8ms, sd = 49.3ms, n = 125
$y = 424.33 + 1.4090x$

Figure 19 - Correlation between deviation score and inter-speaker interval.

Earlier I showed that the first presentation of a map (conversations 1 and 2 throughout the corpus) was accompanied by a higher mean inter-speaker interval than subsequent presentations (conversations 3 to 8 throughout the corpus). This appeared to be caused by a lack of familiarity with the map task and with the sorts of problems encountered in it. Problems that were encountered were therefore not dealt with as efficiently, and so more time was required by the participants to formulate and answer questions, and give and follow appropriate instructions.

An ANOVA comparing conversation number, using deviation score rather than inter-speaker interval as the dependent variable, found that the deviation score of conversations was a significant factor (F (7, 117) = 4.98, p = 0.0001). Further analysis using a Tukey test showed that the mean deviation score for the second conversation in a quad was significantly higher (at the 0.05 level) than the deviation scores for conversations 3, 5, 6, 7, and 8, as shown in Figure 20. This clearly indicates that the less familiar a subject was with the map task, the more problems were encountered (and the greater the mean inter-speaker intervals for the exchanges concerned). These results agree with the Task Familiarity results.

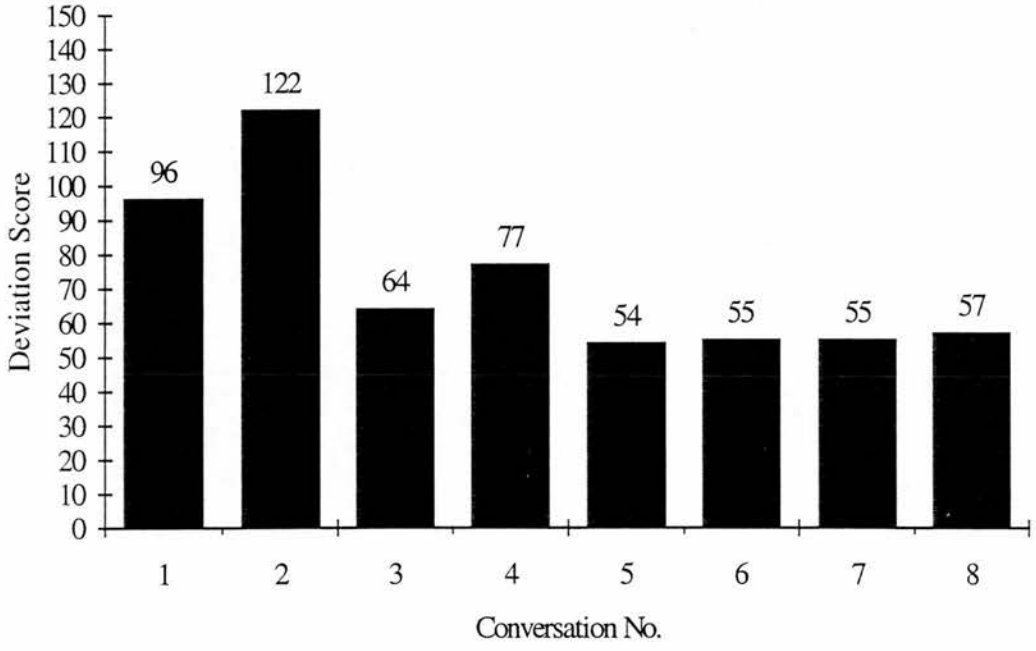


Figure 20 - Mean deviation scores for each conversation number

6.4.9 A Test for the Significance of the Backchannelling Variable with Respect to Inter-Speaker Interval Using a 6-way ANOVA - Game Boundary, Eyecontact, Role, Task Familiarity, A-move channel, and B-move channel

Game Boundary - 2 levels. Boundary, non-boundary.

Eyecontact - 2 levels. Potential for eyecontact, no potential for eye contact.

Role - 2 levels. Giver-follower, follower-giver.

Task Familiarity - 2 levels. Conversation numbers 1 and 2, conversation nos. 3 to 8.

A-move channel - 2 levels. Main channel on the a-move, backchannel on the a-move.

B-move channel - 2 levels. Main channel on the b-move, backchannel on the b-move.

The ANOVA crossing Game Boundary, Eyecontact, Role, Task Familiarity, a-move channel, and b-move channel showed that a-move channel and b-move channels were not significant factors in the determination of inter-speaker interval duration.

A secondary analysis included responses to backchannelling (the backchannel-backchannel, and backchannel-main channel type of exchanges which had been omitted from the original analysis). An ANOVA crossing a-move channel with b-move channel showed that the b-move channel variable was a significant factor in the determination of inter-speaker interval duration (F (1, 13084) = 23.00, $p < 0.0001$), as was the interaction between a-move channel and b-move channel (F (1, 13084) = 26.15, $p < 0.0001$). A Tukey test showed that the significance of the result was due to the relatively high mean inter-speaker interval (671ms) of backchannel/main-channel type exchanges (these were excluded from the initial analysis of backchannelling, for reasons given in Chapter 4). However, the uncertain status of backchannel/main-channel exchanges means that these results have to be treated with some caution. Figure 21 shows the mean inter-speaker intervals of each category.

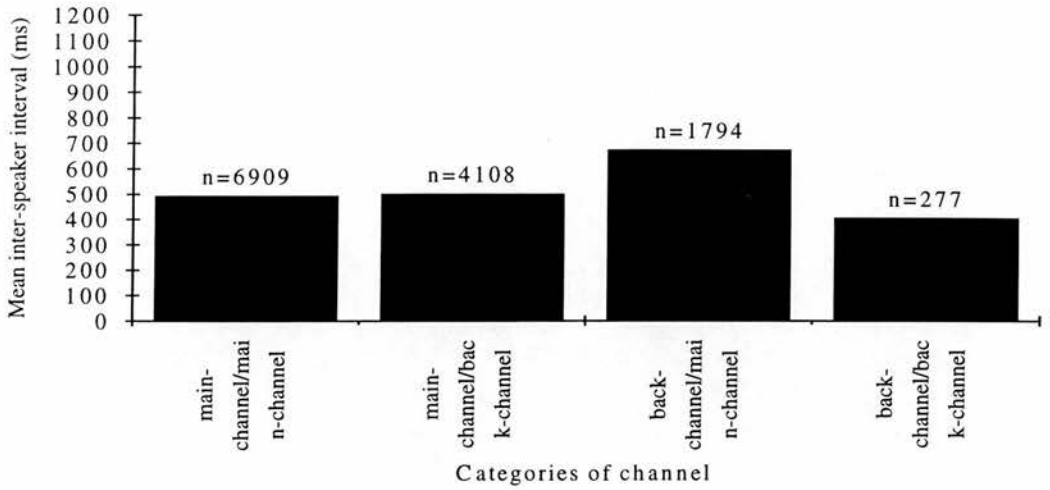


Figure 21 - Mean inter-speaker intervals by a-move channel and b-move channel.

Any backchannelling effect may therefore be caused by the high mean inter-speaker interval of backchannel/mainchannel exchanges, and also possibly an interaction with game boundaries. For example, backchannel/main-channel combinations may have longer inter-speaker intervals because they tend to occur at game boundaries, and not because they are longer *per se*. Figure 22 below shows the means for four channel categories, subcategorised into game internal and game boundary exchanges.

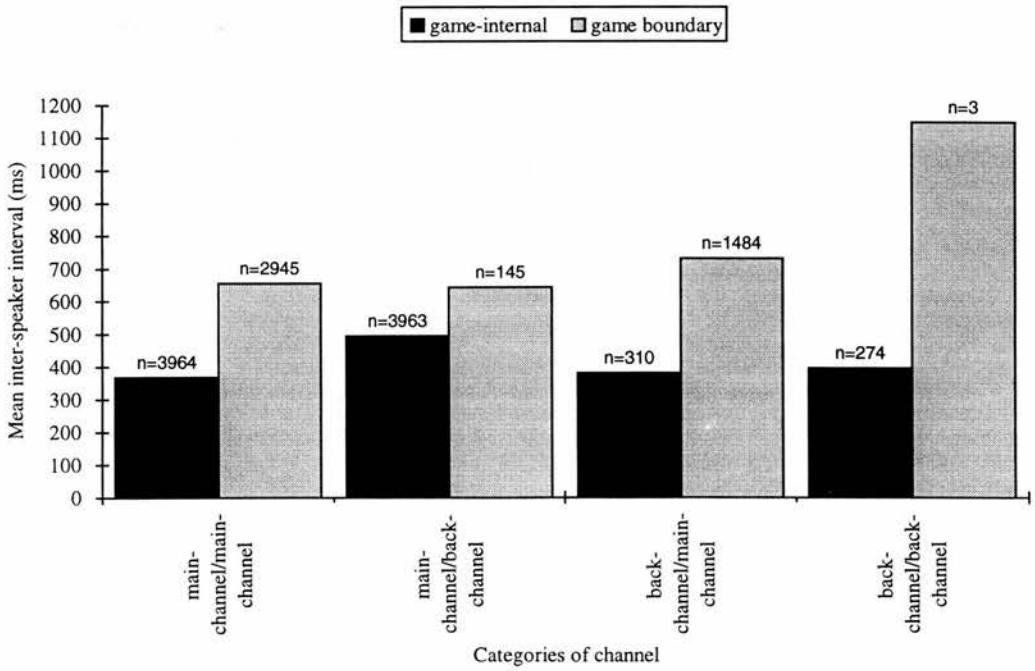


Figure 22 - Mean inter-speaker intervals by channel and game position.

Only a very small percentage of backchannelled utterances occur at the start of a new game (3.38%, as opposed to the 50.89% of mainchannelled utterances which start new games), which may indicate some form of a special status for backchannels. However, some categories of move in the mainchannel were also relatively rare at the starts of games (*reply-y*, *reply-w*, and *reply-n* moves, for example, occurred only 7 times in total at the start of a game). Backchannels may therefore be a category of a more general *REPLY* group.

There are relatively few 'responses' to backchannels game-internally. In other words, the majority of utterances which follow backchannels are mainchannel utterances which start new games. This helps support the case that utterances following backchannels are not generally responses to them.

6.5 Summary and Conclusions[29]

1) One of the most important main effects was for game boundary. Inter-speaker intervals were on average longer at game boundaries than within games, and there was a smaller proportion of overlapping utterances. Since the standard deviation of intervals at boundaries was greater than within games, this suggests that participants were more tolerant of a wider range of positive intervals, and were less likely to overlap or interrupt at boundaries. It is likely that this is the result of the need for more time to plan a new utterance when it follows the end of a previous game.

2) The potential for eye contact was an important main effect, second only to the Game Boundary effect. When partners cannot see each other, inter-speaker intervals are generally shorter than when they can see each other. It was thought that this might have been because the potential to see a partner provides the potential to use facial expression as a coordination cue. Perhaps the presence of this added cue facilitates coordination, and reduces the need for interruptions, or the possible accidental premature start of a contribution. However, the difference between eyecontact and no eyecontact did not lie in the distribution of overlaps. Instead, the difference resulted

---

[29]See Appendix C for a list of significant main effects and interactions.

from a higher standard deviation in the eyecontact cases than in the no eyecontact cases. When participants were able to see each other, there was greater latitude in the range of inter-speaker intervals than when they could not see each other.

3) Mean inter-speaker intervals following an utterance by the information giver were greater than those following an utterance by the information follower. This was contrary to the expectation that giver/follower exchanges should have shorter inter-speaker intervals than follower/giver exchanges should have. The notion that instruction givers require more time to plan their utterances is therefore too simplistic.

Giver/follower exchanges also had a greater standard deviation, and a smaller proportion of overlapping intervals than follower/giver exchanges. The difference between the two cases therefore lies partly in the distribution of positive intervals, which has a broader range with giver/follower exchanges. In the Map Task, instruction followers appear to be under less pressure than givers to respond quickly. This may be because givers have a greater tendency to pass the conversational floor over to followers, than the other way round. If the floor is overtly relinquished to another person, that person would be under less pressure to start as early as possible.

4) Move category was significant. When arranged into a 12-class move system, *check/reply-y*, *explain/acknowledge*, and *query-yn/reply-y* move pairs had inter-speaker intervals which were significantly lower than those of other move pairs. They were also amongst the most common pairs. These may represent 'default' pairs of moves, where minimal thinking time was required by the next speaker to make a response. Two move pairs were found to have intervals significantly longer than other pairs. These were *instruct/ready* and *reply-w/query-w* pairs. These were also relatively uncommon compared to the pairings with shorter mean intervals, and seemed to form less 'natural' classes.

It was also observed that move pairs with longer mean intervals tended to coincide with potential game boundary locations. In fact, an analysis of a 2-way classification of moves into initiators and responses revealed that initiator/initiator

and response/initiator pairs had significantly longer mean intervals than other combinations.

5) Familiarity with a particular task is a significant factor. Typically, the first two conversations in each quad in the Map Task (where neither participant had any experience of the general task) had greater mean inter-speaker intervals than later conversations. The same was true when the conversations were separated into those where both participants had not encountered a certain map before (conversation numbers 1 - 4), and those where one participant (the giver) had encountered one of the maps before (conversation numbers 5 - 8). A better understanding of the task, and the strategies that may or may not work to solve it, therefore led to tighter temporal coordination of turn-taking.

6) The match variable was found to be significant. When there was a match in contrast of the main feature of the giver's and follower's maps, the mean inter-speaker interval was shorter than when there was no match. This result was expected, since no match in the contrast of giver and follower's maps should require more planning time when the non-matching landmarks are encountered, than when the landmarks are the same.

7) There was a small yet significant correlation between the deviation scores and the mean inter-speaker interval for each dialogue. This provided some evidence that the more difficult two participants found a task or problem, the more time was required for some transitions. This effect can be linked also to the level of familiarity with the Map Task, because it was found that the deviation scores in conversations 1 and 2 throughout the corpus were significantly higher than in other conversations.

8) There was no significant difference between the mean inter-speaker intervals preceding backchannelled and main-channelled utterances, except for the case where mainchannel utterances follow backchannels. The status of mainchannel utterances as responses to backchannels is uncertain, however.

This analysis has therefore provided a range of observations of the inter-speaker interval duration in any given exchange, given that enough is known about the context of that exchange (that is, enough variables are known). Important factors in the determination of inter-speaker intervals were game boundaries, move category, familiarity with a given task, and the similarity of the participants' maps. These variables were all concerned with the degree of processing required to make an utterance, or to solve a problem. The assumption here has been that more processing required, the more time taken to plan and produce a response. Both participants are implicitly aware of this relationship, and are able to take account of it in coordinating utterances. I have therefore shown a link between the 'complexity' of an utterance or problem and the following inter-speaker interval. I have also assumed that the notions of earliest possible start and mutual intelligibility also apply as underlying principles. Some variables apply globally to all exchanges in a conversation, such as Eyecontact and Role. The gender of participants, and their degree of familiarity with one another were not found to have a significant influence on inter-speaker intervals.

I have shown that timing depends in part on the relationship between the participants, and what the participants are trying to achieve in their conversation. Also of great importance is the need for time to plan and respond to utterances. The more complex the planning, the more time is needed. I have not been able to consider cultural factors, although one could estimate that they would be highly significant. The interaction of these different factors is complex, and certainly it is difficult to distinguish clearly between them. So, for example, in some circumstances it may not be considered socially desirable, or even necessary, to start an utterance as soon as possible after another person has finished speaking. Instead, there may be a 'pretence' of thinking time. Although this is based on casual observation only, it seems that very often replies to questions such as '*What is your name?*' are not given as soon as possible after the addressee believes he or she knows what the question is and can formulate a reply (which in a suitable contexct may be before the utterance has been finished). Instead, there may be a short, socially acceptable, pause.

7. Conclusions

An important point which I make in this thesis, following from the work of Clark (1996), is that conversation should not be regarded (as it has traditionally been) as a well-ordered system in which each participant takes turns to speak. The structure of real conversation can be highly complex, with as many as 25% of exchanges involving overlap. One participant may reply not to the other participant's last utterance, but to an earlier one. Conversation, according to Clark, should therefore be considered to be a coordination problem, in which participants must coordinate their actions jointly. Each must be able to understand the other, and be able to assume that the other has in turn understood them. This complex process has to be coordinated temporally, by balancing understandability with the need for efficiency. This is not a trivial issue, because the timing of turn-taking has for some time been recognised as a precise process in which slight variations in the intervals between speakers' utterances can be crucial (see for example Couper-Kuhlen, 1993). In this thesis I have been concerned with determining what factors (if any) would play a role in this timing of turn-taking.

One of the main findings of this thesis is that I could find no hard evidence to support the Rhythmic Coordination Hypothesis. In Chapter 5 I described several empirical studies, in which subjects' perceptions of inter-speaker intervals were elicited. They were required to alter the inter-speaker intervals until they appeared to produce a 'natural' length of interval. This operated from the notion that the default case of timing turn-taking was for the next-speaker to time the first prominent syllable in his or her utterance with a perceptual rhythmic beat established in the current speaker's utterance. However, none of these studies revealed any pattern to the distribution of inter-speaker intervals that supported the Rhythmic Coordination Hypothesis.

I also carried out a series of analyses of data from the HCRC Map Task Corpus. Initially, several dialogues from the corpus were hand-labelled for prominent syllables (approximating to pitch accents), and the intervals between these prominences were calculated, both within and across speaker turns. In a second analysis, carried out in conjunction with Matthew Aylett from the HCRC in Edinburgh, the entire Map Task Corpus was automatically labelled for prominences. Both the manually-labelled and automatically-labelled dialogues did not lend any strong support to the Rhythmic Coordination Hypothesis. The expectation was that in a significant proportion of cases, the ratio of the interval between prominences across speaker switches to the interval between prominences within speaker utterances should be equal to an integer.

An alternative hypothesis was therefore tested. This was that the distribution of inter-speaker intervals was influenced not in the main by rhythmic factors, but by two general contextual factors. First, the amount of planning and decision time required to make, understand, and reply to, utterances is important. Second, in conversation people are essentially *doing* something beyond merely conveying the bare semantics of their words. They are attempting to create social relationships with others, establish and modify their common ground, and signal that they have understood each other. They do this within a temporal framework, which allows for the management of the scarce resource.

The Map Task Corpus provided a useful means of linking these two broad categories to inter-speaker interval duration. However, in the final analysis it was not possible to distinguish neatly between variables which are concerned with planning and decision time ('cognitive' variables) and those concerned with what the participants are doing ('communicative' variables), since this is an idealised abstraction.

I carried out an analysis of the mean inter-speaker intervals of several variables which had previously been coded in the Map Task Corpus. These were:

i) whether eye contact was possible in a conversation (Eyecontact),

ii) whether an inter-speaker interval was followed by a giver or follower's utterance (Role),

iii) whether an interval was followed by a conversational game boundary (Game Boundary),

iv) whether there was a contrast in the names of the two master features of the instruction giver's map (Contrast),

v) whether there were differences in the master feature landmarks on the giver and follower's maps (Match),

vi) the route design of the map used in a dialogue,

vii) the move category of the last move before an inter-speaker interval (A-Move), and the move category of the first move after an inter-speaker interval (B-Move),

viii) whether the participants had encountered the Map Task before (Task Familiarity),

ix) the gender of the participants (Sex),

x) whether the participants knew each other or not (Familiarity),

xi) whether the two participants were actually looking up or down (Gaze), as opposed to the *potential* for eye contact as measured by the Eyecontact variable,

xii) whether there was first mention of a shared or unshared landmark in any given exchange (Shared Landmark),

xiii) the extent to which the route drawn by the follower differed from the route marked on the giver's map (Deviation Score),

xiv) whether an exchange consisted of backchannelling or a main channel utterance in either the first or second turn (Backchannelling).

Of these, Eyecontact, Role, Game Boundary, Match, Task Familiarity, A-Move, B-Move, and Deviation Score were significant.

In the case of Eyecontact, the mean inter-speaker interval when both participants could see each other was significantly greater than when they could not. One might have expected the reverse to be true, since not seeing the person one is talking to could produce difficulties not present in face-to-face interaction. This would necessitate greater planning and decision time, and mean inter-speaker

intervals would therefore be greater. The explanation for the actual results is that the potential for a listener to see the speaker presents less urgency for the listener to interrupt or overlap when he or she wishes to start speaking. There is a greater tolerance of a broader range of positive inter-speaker intervals when people can see each other than when they cannot.

When exchanges were of the giver-follower type the inter-speaker intervals were significantly longer than when they were of the follower-giver type. That is, instruction followers left a longer mean interval before starting their utterances than did instruction givers. This may well be the result of followers being 'handed' the conversational floor more frequently than givers were. If one speaker gives instructions, or asks a question, some form of direct response to that may be expected. The conversational floor has been explicitly offered to the other participant, and although the previous speaker may start a new utterance at any time, there is a tolerance of inter-speaker intervals that are longer than would be encountered when the floor has not actively been offered.

The other significant variables - Game Boundary, Move Category, Match, Task Familiarity, and Deviation Score - are related directly to the planning/inter-speaker interval relationship. For example, one might reasonably expect that at game boundaries more planning would be required, and so the inter-speaker interval would be greater, than in exchanges which are not at game boundaries. This was found to be so. Also, as the difficulty of the overall task, or of parts of it, increased, mean inter-speaker intervals should increase. Again, results from the Match, Task Familiarity, and Deviation Score main effects support this notion. When participants were unfamiliar with the overall task (as in dialogues 1 and 2), mean inter-speaker intervals were greater than when they were (as in dialogues 7 and 8). Also, there was some correlation between the deviation score for each dialogue's map, and the mean inter-speaker interval for that dialogue. Difficulties with the overall task involved in a dialogue therefore altered the mean intervals.

Particularly striking were the results for the Match variable. Where giver and follower's maps did not have matching contrast in their master features, the mean inter-speaker interval was greater than where there was a match. Therefore, the

difference in only one feature between giver and follower's maps was enough to cause the mean interval for that dialogue to be greater than when there was no such difference.

The results from the move coding analysis also proved insightful. Again, these results fit in generally with the planning/inter-speaker interval relationship. However, exact predictions were difficult in this case because it was difficult to know whether any given category of move would require inherently more time to produce or understand than any other category. Amongst other observations, I found that *reply-y* moves tended to follow relatively quickly after a previous speaker's utterance, and that *reply-y* moves very often followed *align*, *check*, and *query-yn* moves. On the other hand, *instruct* and *ready* moves tended to be followed by relatively long inter-speaker intervals, and *instruct* moves were very often followed by *query-w* and *ready* moves. Generally, the move coding analysis highlighted the importance of semantic considerations in the timing of turn-taking. As noted earlier, one of the findings of this thesis is that is that the move coding system used in the Map Task Corpus (and indeed in the discourse literature in general) is often too gross. Certainly this was a problem with the move coding analysis.

A point needs to be made about backchannelling. A perhaps surprising result of this research was that there was on average no significant difference between inter-speaker intervals preceding backchannel and mainchannel utterances. This was not expected, because backchannel utterances are conventionally considered to be placed anywhere with respect to the other speaker's mainchannel utterance. In other words, traditional accounts consider backchannelling to occur outwith the main channel of conversation, and therefore outwith the normal rules of turn-taking timing and coordination. In fact, one might normally think of backchannelled utterances as overlapping with mainchannel utterances more than main channel utterances would overlap with each other. The mean inter-speaker interval would therefore be smaller. Backchannelling therefore seems to be treated, at least temporally, in the same way as mainchannelling by interlocutors.

While backchannelling does have different structural and functional roles from many forms of mainchannel (such as its rarity at the starts of game boundaries),

it does not differ from all forms. It is not therefore possible to make a clear backchannel/mainchannel distinction. This is highlighted by the difficulty of finding an adequate definition of backchannelling to begin with. In this thesis, the basic definition of a backchannel has been that it is an *acknowledge* move. But this is almost certainly too cumbersome a definition, and would require revision in further research.

A further problem with backchannelling, even where it is clearly defined, is that a clear measurement of inter-speaker intervals is fraught with difficulties. What might a given backchannel be a response to, and hence what inter-speaker interval should be measured? Is it possible for someone to respond to a backchannel directly? And if not, then would the structure of the conversation break down without backchannelling? Backchannelling may serve the function of a signal which says 'yes, I'm still listening, so carry on talking'. Therefore, without it the conversation may falter.

Why, then, bother at all with backchannelling as a separate entity, particularly when attempts to do so seem to create more problems than they solve? Indeed, the arguments against such a treatment are convincing. Backchannelling may be thought of perhaps as a sub-category of the *reply-y* move. It has essentially the same function, since it acts to give a positive feedback, and very often the two categories have the same structure - '*mmhmm*' seems on preliminary inspection to be common amongst both categories. The difference lies in the nature of the utterance which the backchannel or *reply-y* responds to. A *reply-y* move will generally respond to an utterance which requires a direct response, whereas a backchannel seems to be used voluntarily. The distinction between mainchannel and backchannel is therefore a useful one, and what difference there are seem to lie in the nature of what each *does* in terms of communication.

Much of Chapter 3 was concerned with trying to form rules which could determine whether an utterance could be considered to be a response to another utterance, and if so, which utterance it could be a response to. I formulated a part-solution in terms of inter-speaker intervals. But there is as yet no clear solution to this problem. The answer must surely lie to a large extent in a more detailed and thorough

analysis of the semantics involved in an exchange. The move coding used in this analysis gave some semantic information, although as I mentioned above it was too broadly defined to provide the full range of subtlety required for a temporal analysis. For example, not all wh-questions are equally complex, and therefore would not be expected to be preceded by similar inter-speaker intervals.

I also mentioned in Chapter 3 the need for a second form of revision to move coding in general. The uncertainties over what constituted a response to what highlighted the need for units that are smaller than moves - what may be termed *sub-move units*, but which in fact approximate closer than moves to the *turn constructional unit* (or TCU, see Chapter 2). Very often, it seemed that one move was responding to an earlier part of the other speaker's move. The *function* of both parts may be similar, but in terms of the real-time interaction of two participants, the move must be thought of as consisting of two parts, separated by a TCU boundary - a transition relevance place (TRP). This research has therefore further highlighted the importance of the TCU and TRP as real entities in conversation, which deserve further study if the coordination of turn-taking is to be understood better.

In Chapter 2, I have suggested a series of components of a turn-taking coordination model. It is reiterated here:

1) Coordination component

This covers the need for conversation to be coordinated by some means - either through a rhythmic system, or through a linear system. In this thesis, I have favoured a linear system, because I could find no evidence in favour of a rhythmic coordination model.

2) Communicative component

Conversation is largely concerned with what the participants are doing, and how they are interacting. A model of the timing of turn-taking (and of turn-taking in general) must therefore account for the relationships between people, and between people and their environment and context. This thesis has found no direct evidence for this component, although this is largely because of the difficulties in isolating variables

which are concerned purely with how the participants are interacting, what their relationship to each other is.

3) Cognitive component

According to the principle of processing time (see Clark, 1996), different utterances will require different amounts of time to plan and respond to, depending on the complexity of the utterance. This thesis has shown that inter-speaker intervals are related to the requirements that planning and decision place on conversation.

4) Projection component

The projection of an entry point requires projection of a likely closure of C's contribution. This is most probably achieved through the projection of TRPs. This is a vital issue in understanding the timing and coordination of turn-taking, but the exact means by which projection is achieved are not certain. Although I did not look directly into this area, I did find evidence that units smaller than moves appear to be used in the coordination of turn-taking. It seems quite plausible that these sub-units are equivalent to the turn constructional units hypothesised by, for example, Ford & Thompson (1995) as the basic units used in projection.

Appendix A

A1. Representations in SGML Data Files

Words

In the data files, each new word was represented by a line of code, bounded by the 'W' marker. In each line was the start time in seconds (marked by START=n), its duration in seconds (marked by DUR=n), and part of speech (TAG=x). The word itself was listed between angled brackets. An example line of code for the word 'starting' is:

1)

<W START=1.8394 DUR=0.5111 TAG=vbg>starting</W>

Some 'words' were listed as figures in square brackets. For example:

2)

<W START=0.0000 DUR=8.2034 TAG=sent>[8.2034]</W>

This figure represents a silence by that speaker, in seconds.

Moves

The start and end of each move was signalled by a 'move' marker. The code for each move consisted of a move number marker (id=n), and a move type marker (label=x). All speech in a move was included within the move boundary markers, as shown below:

3)

<move id=m3 label=acknowledge>
<W START=0.0000 DUR=8.2034 TAG=sent>[8.2034]</W>
<W START=8.2034 DUR=0.3375 TAG=aff>mmhmm</W>
</move>

Turns

Conversational turns were marked in the SGML code, with their number marker (id=n), and whether they were uttered by the information giver or follower (who=x). Turns were bounded by the 'turn' marker.

4)

<turn id=t1 who=giver>

Games

The start of a new game is marked by the 'startgame' marker. Listed are the game identification marker (id=x), whether the person starting the game was the giver or follower (INITIATOR = x), and the game category (type = x). If a game was embedded, an extra marker was used (em = true-em).

5)

<startgame primaryview="actions" id=acg2 INITIATOR = giver type = explain em = true-em>

The end of a game is marked by the 'endgame' marker. Listed in the endgame line is the start-id of that game (start-id=x).

6)

<endgame primaryview="actions" start-id=acg2>

Gaze

Text is an alternating series of *lookon's* and *lookoff's*, starting with *lookon* and ending in either. A *lookon* is divided into a *startlookup* and *startlookdown*. The end of a *startlookup* is marked by an *endlookup*, which must be accompanied by a *startlookdown*. Likewise, an *endlookon* must be followed by a *startlookoff*. Included in the gaze coding is information on which speaker is involved in the change of gaze status (e.g. primaryview="followergaze"). There is also an identification marker (id=x). Examples of gaze coding follow.

The code in 7) shows that the information follower started to look on (foll0), and then immediately looked down (foll1).

7)

<startlookon primaryview="followergaze" id=foll0>

<startlookdown primaryview="followergaze" id=foll1>

In 8), the information follower looked up (the start of foz0). After a noise and a short silence (indicated by the two lines beginning with '<W'), the follower stopped looking up (the end of foz0), and looked down (the start of foll4).

8)

<startlookup primaryview="followergaze" id=foz0>

<W START=0.0000 DUR=0.5751 TYPE=UNINTELLIGIBLE TAG=noi>#x</W>

<W START=0.5751 DUR=1.2643 TAG=sent>[1.2643]</W>

<endlookup primaryview="followergaze" start-id=foz0>

<startlookdown primaryview="followergaze" id=foll4>

In 9), the follower stopped looking on (the end of the lookon marker foll0), and started to look off (fol1).

9)

<endlookon primaryview="followergaze" start-id=foll0>

<startlookoff primaryview="followergaze" id=fol1>

10) shows that the follower stopped looking off (the end of the lookoff marker foll3).

10)

<endlookoff primaryview="followergaze" start-id=fol13>

A2. A fragment from a data file, written in SGML format.

```
<!doctype text system "maptask-turns.dtd" >
<text id=q3ec1>
<!-- Conversation: Quad 3, eye contact, conversation 1, -->
<!--               unfamiliar talkers, duration 10606393 -->
<!--        Giver: q3et4141 -->
<!--                Philip, age 18, birth.place Glasgow, male -->
<!--     Follower: q3et4242 -->
<!--                Ross, age 20, birth.place Glasgow, male -->
<!--          Map: m14, +giver contrast, +follower match,
reduction type 3 -->
<!-- Copyright 1992, Human Communication Research Centre -->
<startgame primaryview="actions" id=acg1 INITIATOR = giver
type=uncoded>
<turn  id=t1 who=giver>
<move id=m1 label=instruct>
<startlookon primaryview="followergaze" id=foll0>
<startlookdown primaryview="followergaze" id=foll1>
<startlookon primaryview="givergaze" id=gill2>
<startlookdown primaryview="givergaze" id=gill3>
<timestamp time=350>
<startlookup primaryview="followergaze" id=foz0>
<W START=0.0000 DUR=0.5751 TYPE=UNINTELLIGIBLE TAG=noi>#x</W>
<W START=0.5751 DUR=1.2643 TAG=sent>[1.2643]</W>
<endlookup primaryview="followergaze" start-id=foz0>
<startlookdown primaryview="followergaze" id=foll4>
<W START=1.8394 DUR=0.5111 TAG=vbg>starting</W>
<W START=2.3505 DUR=0.2374 TYPE=OUTBREATH TAG=noi>#ho</W>
<W START=2.5879 DUR=0.2058 TAG=unknown>[0.2058]</W>
<W START=2.7937 DUR=0.0602 TAG=in>at</W>
<startlookup primaryview="givergaze" id=giz1>
<W START=2.8539 DUR=0.0392 TAG=at>the</W>
<W START=2.8931 DUR=0.3386 TAG=nn>beginning</W>
<endlookup primaryview="givergaze" start-id=giz1>
<startlookdown primaryview="givergaze" id=gill5>
<W TAG=cm>,</W>
<W START=3.2317 DUR=0.5325 TAG=unknown>[0.5325]</W>
<W START=3.7643 DUR=0.2780 TAG=vb>head</W>
<startlookup primaryview="givergaze" id=giz2>
<W START=4.0423 DUR=0.1925 TAG=ql>due</W>
<W START=4.2348 DUR=0.4458 TAG=rp>south</W>
<endlookup primaryview="givergaze" start-id=giz2>
<startlookdown primaryview="givergaze" id=gill6>
<W TAG=sent>.</W>
<W START=4.6806 DUR=1.2191 TAG=unknown>[1.2191]</W>
<ucode ucontent="startjj same intro men nl in dir">
<scode scontent="topic brsyn partphr mc imp adv direct">
</move>
<startgame primaryview="actions" id=acg2 INITIATOR = giver type =
explain em = true-em>
<move id=m2 label=explain>
<endlookon primaryview="followergaze" start-id=foll0>
<startlookoff primaryview="followergaze" id=fol1>
<W START=5.8996 DUR=0.1166 TAG=ppss>you</W>
<W TYPE=CLITIC TAG=md>+'ll</W>
<W START=6.0163 DUR=0.2290 TAG=vb>see</W>
<W START=6.2453 DUR=0.0678 TAG=at>a</W>
<W START=6.3131 DUR=0.5839 TAG=nn>diamond_mine</W>
```

```
<startlookup primaryview="givergaze" id=giz3>
<W TAG=cm>,</W>
<W START=6.8970 DUR=0.2560 TYPE=OUTBREATH TAG=noi>#ho</W>
<W START=7.1530 DUR=0.2067 TAG=unknown>[0.2067]</W>
<W START=7.3598 DUR=0.0991 TAG=in>on</W>
<W START=7.4589 DUR=0.0460 TAG=ppg>your</W>
<endlookup primaryview="givergaze" start-id=giz3>
<startlookdown primaryview="givergaze" id=gil17>
<W START=7.5049 DUR=0.3801 TAG=nn>map</W>
<ucode ucontent="dmine ddel contrsh same intro men indef 1 e">
<scode scontent="partphr mc dec mod adv locat pp">
</move>
<turn  id=t2 who=follower>
<move id=m3 label=acknowledge>
<endlookoff primaryview="followergaze" start-id=fol1>
<startlookon primaryview="followergaze" id=foll8>
<startlookdown primaryview="followergaze" id=foll9>
<timestamp time=160913>
<W START=0.0000 DUR=8.2034 TAG=sent>[8.2034]</W>
<W START=8.2034 DUR=0.3375 TAG=aff>mmhmm</W>
<W TAG=sent>.</W>
<ucode ucontent="resp +">
</move>
<endgame primaryview="actions" start-id=acg2>
```

Appendix B

B1.

An example of the cell sizes of a-move crossed with b-move, when all possible categories of move are used. Note that in many cases, the cells either have no members, or only very few.

a_move	b_move	COUNT	a_move	b_move	COUNT
alig	ackn	56	expl	clar	9
alig	alig	6	expl	expl	76
alig	chec	84	expl	inst	45
alig	clar	1	expl	read	130
alig	expl	25	expl	repn	39
alig	inst	2	expl	repy	22
alig	read	18	expl	repw	1
alig	repn	37	expl	quyn	47
alig	repy	730	expl	qu_w	44
alig	repw	26	inst	ackn	1578
alig	quyn	15	inst	alig	8
alig	qu_w	14	inst	chec	508
chec	ackn	8	inst	clar	0
chec	alig	13	inst	expl	112
chec	chec	14	inst	inst	5
chec	clar	152	inst	read	98
chec	expl	28	inst	repn	10
chec	inst	6	inst	repy	10
chec	read	41	inst	repw	1
chec	repn	174	inst	quyn	134
chec	repy	868	inst	qu_w	144
chec	repw	46	read	ackn	18
chec	quyn	17	read	alig	4
chec	qu_w	17	read	chec	8
clar	ackn	427	read	clar	1
clar	alig	2	read	expl	6
clar	chec	105	read	inst	10
clar	clar	0	read	read	6
clar	expl	39	read	repn	0
clar	inst	3	read	repy	1
clar	read	22	read	repw	0
clar	repn	9	read	quyn	6
clar	repy	10	read	qu_w	2
clar	repw	2	repn	ackn	180
clar	quyn	15	repn	alig	4
clar	qu_w	20	repn	chec	35
expl	ackn	623	repn	clar	1
expl	alig	24	repn	expl	32
expl	chec	101	repn	inst	9

a_move	b_move	COUNT
repn	read	37
repn	repn	0
repn	repy	1
repn	repw	0
repn	quyn	39
repn	qu_w	11
repy	ackn	469
repy	alig	61
repy	chec	96
repy	clar	12
repy	expl	51
repy	inst	192
repy	read	285
repy	repn	3
repy	repy	21
repy	repw	9
repy	quyn	67
repy	qu_w	47
repw	ackn	336
repw	alig	6
repw	chec	65
repw	clar	3
repw	expl	40
repw	inst	11
repw	read	53

a_move	b_move	COUNT
repw	repn	6
repw	repy	5
repw	repw	3
repw	quyn	25
repw	qu_w	26
quyn	ackn	21
quyn	alig	5
quyn	chec	28
quyn	clar	8
quyn	expl	31
quyn	inst	6
quyn	read	18
quyn	repn	410
quyn	repy	640
quyn	repw	137
quyn	quyn	14
quyn	qu_w	18
qu_w	ackn	10
qu_w	alig	15
qu_w	chec	10
qu_w	clar	70
qu_w	expl	11
qu_w	inst	4
qu_w	read	53
qu_w	repn	11
qu_w	repy	8
qu_w	repw	346
qu_w	quyn	11
qu_w	qu_w	8

B2.

An example of the cell sizes, and mean inter-speaker intervals of a-move/b-move pairings, where 6 categories of a-move and 9 categories of b-move were used. Note that in many cases, the cells either have no members, or only very few.

CELL MEANS

a_move =	chec	chec	chec	chec	chec
b_move =	ackn	alig	chec	expl	inst
interval	491.33750	231.14615	890.73572	636.12500	131.98334
COUNT	8	13	14	28	6

a_move =	chec	chec	chec	chec	expl
b_move =	read	repy	quyn	qu_w	ackn
interval	652.40976	176.31982	902.22943	768.80588	365.83178
COUNT	41	868	17	17	623

a_move =	expl	expl	expl	expl	expl
b_move =	alig	chec	expl	inst	read
interval	908.77082	619.60594	643.03948	370.58889	610.72385
COUNT	24	101	76	45	130

a_move =	expl	expl	expl	inst	inst
b_move =	repy	quyn	qu_w	ackn	alig
interval	388.92273	504.85106	773.80683	632.62934	671.67499
COUNT	22	47	44	1578	8

a_move =	inst	inst	inst	inst	inst
b_move =	chec	expl	inst	read	repy
interval	660.27598	740.10625	908.48001	938.94592	83.48000
COUNT	508	112	5	98	10

a_move =	inst	inst	repy	repy	repy
b_move =	quyn	qu_w	ackn	alig	chec
interval	771.56045	800.14444	533.69573	601.61639	506.38229
COUNT	134	144	469	61	96

a_move =	repy	repy	repy	repy	repy
b_move =	expl	inst	read	repy	quyn
interval	458.41176	582.60677	540.28702	178.02857	542.83731
COUNT	51	192	285	21	67

a_move =	repy	repw	repw	repw	repw
b_move =	qu_w	ackn	alig	chec	expl
interval	901.26595	460.25327	1408.73332	563.36923	597.01001
COUNT	47	336	6	65	40

a_move =	repw	repw	repw	repw	repw
b_move =	inst	read	repy	quyn	qu_w
interval	887.47272	698.21699	1137.06000	612.50000	1166.14999
COUNT	11	53	5	25	26

a_move =	quyn	quyn	quyn	quyn	quyn
b_move =	ackn	alig	chec	expl	inst
interval	330.03810	319.58000	847.43214	566.04516	210.05000
COUNT	21	5	28	31	6

a_move =	quyn	quyn	quyn	quyn	MARGINAL
b_move =	read	repy	quyn	qu_w	
interval	895.99999	328.18375	427.08571	629.20555	520.69698
COUNT	18	640	14	18	7358

B3.

An example of the cell sizes of a 9-way ANOVA, crossing game boundary, eyecontact, role, familiarity, sex, task-familiarity, match, route, and contrast. Even though more variables were available, the cell sizes are either low, or zero. Clearly, it was not possible to test all the variables in the analysis in a single n-way ANOVA.

gameboun	eye	role	fam	sex	taskfam	match	route	contrast	COUNT
*0	noeye	*1	unfam	*1	*0	*0	*1	*0	57
*0	noeye	*1	unfam	*1	*0	*0	*1	*1	13
*0	noeye	*1	unfam	*1	*0	*0	*3	*0	16
*0	noeye	*1	unfam	*1	*0	*0	*4	*0	12
*0	noeye	*1	unfam	*1	*0	*1	*1	*0	11
*0	noeye	*1	unfam	*1	*0	*1	*2	*0	33
*0	noeye	*1	unfam	*1	*0	*1	*2	*1	16
*0	noeye	*1	unfam	*1	*0	*1	*4	*0	60
*0	noeye	*1	unfam	*1	*0	*1	*4	*1	24
*0	noeye	*1	unfam	*1	*1	*0	*3	*1	37
*0	noeye	*1	unfam	*1	*1	*1	*1	*1	95
*0	noeye	*1	unfam	*1	*1	*1	*2	*1	27
*0	noeye	*1	unfam	*1	*1	*1	*3	*1	17
*0	noeye	*1	unfam	*2	*0	*0	*2	*0	29
*0	noeye	*1	unfam	*2	*0	*0	*2	*1	8
*0	noeye	*1	unfam	*2	*0	*0	*3	*1	35
*0	noeye	*1	unfam	*2	*0	*0	*4	*0	23
*0	noeye	*1	unfam	*2	*0	*0	*4	*1	12
*0	noeye	*1	unfam	*2	*0	*1	*1	*0	25
*0	noeye	*1	unfam	*2	*0	*1	*1	*1	16
*0	noeye	*1	unfam	*2	*0	*1	*3	*0	34
*0	noeye	*1	unfam	*2	*0	*1	*3	*1	26
*0	noeye	*1	unfam	*2	*0	*1	*4	*0	4
*0	noeye	*1	unfam	*2	*1	*0	*2	*1	54
*0	noeye	*1	unfam	*2	*1	*0	*4	*1	15
*0	noeye	*1	unfam	*3	*0	*0	*3	*0	9
*0	noeye	*1	unfam	*3	*0	*1	*2	*0	21
*0	noeye	*1	unfam	*3	*1	*0	*1	*1	33
*0	noeye	*1	unfam	*3	*1	*1	*4	*1	17
*0	noeye	*1	fam	*1	*0	*0	*1	*0	4
*0	noeye	*1	fam	*1	*0	*0	*2	*1	33
*0	noeye	*1	fam	*1	*0	*0	*4	*0	21
*0	noeye	*1	fam	*1	*0	*0	*4	*1	21
*0	noeye	*1	fam	*1	*0	*1	*1	*0	26
*0	noeye	*1	fam	*1	*1	*0	*3	*1	2
*0	noeye	*1	fam	*2	*0	*0	*2	*0	40
*0	noeye	*1	fam	*2	*0	*0	*4	*0	17
*0	noeye	*1	fam	*2	*0	*1	*1	*0	8
*0	noeye	*1	fam	*2	*0	*1	*1	*1	46
*0	noeye	*1	fam	*2	*0	*1	*2	*0	21
*0	noeye	*1	fam	*2	*0	*1	*2	*1	41
*0	noeye	*1	fam	*2	*0	*1	*3	*0	72
*0	noeye	*1	fam	*2	*0	*1	*3	*1	15
*0	noeye	*1	fam	*2	*0	*1	*4	*0	112
*0	noeye	*1	fam	*2	*1	*0	*2	*1	25
*0	noeye	*1	fam	*2	*1	*0	*4	*1	51
*0	noeye	*1	fam	*2	*1	*1	*1	*1	21
*0	noeye	*1	fam	*2	*1	*1	*2	*1	19
*0	noeye	*1	fam	*2	*1	*1	*3	*1	27
*0	noeye	*1	fam	*2	*1	*1	*4	*1	12
*0	noeye	*1	fam	*3	*0	*0	*1	*0	35
*0	noeye	*1	fam	*3	*0	*0	*1	*1	28
*0	noeye	*1	fam	*3	*0	*0	*3	*0	49
*0	noeye	*1	fam	*3	*0	*0	*3	*1	34
*0	noeye	*1	fam	*3	*0	*1	*2	*0	34
*0	noeye	*1	fam	*3	*0	*1	*4	*1	43
*0	noeye	*1	fam	*3	*1	*0	*1	*1	62
*0	noeye	*2	unfam	*1	*0	*0	*1	*0	91
*0	noeye	*2	unfam	*1	*0	*0	*1	*1	43
*0	noeye	*2	unfam	*1	*0	*0	*3	*0	53
*0	noeye	*2	unfam	*1	*0	*0	*4	*0	35
*0	noeye	*2	unfam	*1	*0	*1	*1	*0	25
*0	noeye	*2	unfam	*1	*0	*1	*2	*0	27
*0	noeye	*2	unfam	*1	*0	*1	*2	*1	39
*0	noeye	*2	unfam	*1	*0	*1	*4	*0	56
*0	noeye	*2	unfam	*1	*0	*1	*4	*1	39
*0	noeye	*2	unfam	*1	*1	*0	*3	*1	50
*0	noeye	*2	unfam	*1	*1	*1	*1	*1	95
*0	noeye	*2	unfam	*1	*1	*1	*2	*1	51
*0	noeye	*2	unfam	*1	*1	*1	*3	*1	29
*0	noeye	*2	unfam	*2	*0	*0	*2	*0	62
*0	noeye	*2	unfam	*2	*0	*0	*2	*1	35
*0	noeye	*2	unfam	*2	*0	*0	*3	*1	32
*0	noeye	*2	unfam	*2	*0	*0	*4	*0	51
*0	noeye	*2	unfam	*2	*0	*0	*4	*1	18
*0	noeye	*2	unfam	*2	*0	*1	*1	*0	29

*0	noeye	*2	unfam	*2	*0	*1	*1	*1	36
*0	noeye	*2	unfam	*2	*0	*1	*3	*0	74
*0	noeye	*2	unfam	*2	*0	*1	*3	*1	34
*0	noeye	*2	unfam	*2	*0	*1	*4	*0	33
*0	noeye	*2	unfam	*2	*1	*0	*2	*1	52
*0	noeye	*2	unfam	*2	*1	*0	*4	*1	33
*0	noeye	*2	unfam	*3	*0	*0	*3	*0	32
*0	noeye	*2	unfam	*3	*0	*1	*2	*0	50
*0	noeye	*2	unfam	*3	*1	*0	*1	*1	72
*0	noeye	*2	unfam	*3	*1	*1	*4	*1	20
*0	noeye	*2	fam	*1	*0	*0	*1	*0	33
*0	noeye	*2	fam	*1	*0	*0	*2	*1	44
*0	noeye	*2	fam	*1	*0	*0	*4	*0	41
*0	noeye	*2	fam	*1	*0	*0	*4	*1	29
*0	noeye	*2	fam	*1	*0	*1	*1	*0	31
*0	noeye	*2	fam	*1	*1	*0	*3	*1	7
*0	noeye	*2	fam	*2	*0	*0	*2	*0	65
*0	noeye	*2	fam	*2	*0	*0	*4	*0	33
*0	noeye	*2	fam	*2	*0	*1	*1	*0	30
*0	noeye	*2	fam	*2	*0	*1	*1	*1	70
*0	noeye	*2	fam	*2	*0	*1	*2	*0	37
*0	noeye	*2	fam	*2	*0	*1	*2	*1	100
*0	noeye	*2	fam	*2	*0	*1	*3	*0	79
*0	noeye	*2	fam	*2	*0	*1	*3	*1	18
*0	noeye	*2	fam	*2	*0	*1	*4	*0	138
*0	noeye	*2	fam	*2	*1	*0	*2	*1	38
*0	noeye	*2	fam	*2	*1	*0	*4	*1	63
*0	noeye	*2	fam	*2	*1	*1	*1	*1	41
*0	noeye	*2	fam	*2	*1	*1	*2	*1	37
*0	noeye	*2	fam	*2	*1	*1	*3	*1	30
*0	noeye	*2	fam	*2	*1	*1	*4	*1	32
*0	noeye	*2	fam	*3	*0	*0	*1	*0	58
*0	noeye	*2	fam	*3	*0	*0	*1	*1	51
*0	noeye	*2	fam	*3	*0	*0	*3	*0	63
*0	noeye	*2	fam	*3	*0	*0	*3	*1	46
*0	noeye	*2	fam	*3	*0	*1	*2	*0	70
*0	noeye	*2	fam	*3	*0	*1	*4	*1	29
*0	noeye	*2	fam	*3	*1	*0	*1	*1	61
*0	eye	*1	unfam	*1	*0	*0	*1	*0	16
*0	eye	*1	unfam	*1	*0	*0	*1	*1	23
*0	eye	*1	unfam	*1	*0	*0	*3	*0	16
*0	eye	*1	unfam	*1	*0	*0	*4	*0	23
*0	eye	*1	unfam	*1	*0	*1	*1	*1	8
*0	eye	*1	unfam	*1	*0	*1	*2	*0	49
*0	eye	*1	unfam	*1	*0	*1	*4	*0	26
*0	eye	*1	unfam	*1	*1	*0	*1	*1	14
*0	eye	*1	unfam	*1	*1	*0	*3	*1	32
*0	eye	*1	unfam	*1	*1	*1	*1	*1	4
*0	eye	*1	unfam	*1	*1	*1	*2	*1	11
*0	eye	*1	unfam	*1	*1	*1	*4	*1	25
*0	eye	*1	unfam	*2	*0	*0	*3	*0	23
*0	eye	*1	unfam	*2	*0	*1	*3	*0	12
*0	eye	*1	unfam	*2	*0	*1	*4	*1	15
*0	eye	*1	unfam	*2	*1	*0	*2	*1	14
*0	eye	*1	unfam	*3	*0	*0	*1	*0	33
*0	eye	*1	unfam	*3	*0	*0	*2	*0	44
*0	eye	*1	unfam	*3	*0	*0	*2	*1	19
*0	eye	*1	unfam	*3	*0	*0	*4	*1	33
*0	eye	*1	unfam	*3	*0	*1	*1	*0	19
*0	eye	*1	unfam	*3	*0	*1	*2	*1	3
*0	eye	*1	unfam	*3	*0	*1	*3	*0	16
*0	eye	*1	unfam	*3	*0	*1	*3	*1	16
*0	eye	*1	unfam	*3	*0	*1	*4	*0	23
*0	eye	*1	unfam	*3	*1	*0	*4	*1	5
*0	eye	*1	unfam	*3	*1	*1	*3	*1	30
*0	eye	*1	fam	*1	*0	*0	*3	*0	14
*0	eye	*1	fam	*1	*0	*0	*4	*0	16
*0	eye	*1	fam	*1	*0	*1	*1	*1	17
*0	eye	*1	fam	*1	*0	*1	*3	*0	31
*0	eye	*1	fam	*1	*1	*0	*1	*1	55
*0	eye	*1	fam	*1	*1	*0	*2	*1	26
*0	eye	*1	fam	*1	*1	*1	*2	*1	38
*0	eye	*1	fam	*1	*1	*1	*3	*1	18
*0	eye	*1	fam	*2	*0	*0	*2	*1	12
*0	eye	*1	fam	*2	*0	*0	*4	*0	9
*0	eye	*1	fam	*2	*0	*1	*2	*0	34
*0	eye	*1	fam	*2	*0	*1	*2	*1	31
*0	eye	*1	fam	*2	*0	*1	*4	*0	37
*0	eye	*1	fam	*2	*0	*1	*4	*1	12
*0	eye	*1	fam	*2	*1	*1	*4	*1	28
*0	eye	*1	fam	*3	*0	*0	*1	*0	92
*0	eye	*1	fam	*3	*0	*0	*1	*1	19
*0	eye	*1	fam	*3	*0	*0	*2	*0	13
*0	eye	*1	fam	*3	*0	*0	*3	*0	32
*0	eye	*1	fam	*3	*0	*0	*3	*1	26
*0	eye	*1	fam	*3	*0	*0	*4	*1	18
*0	eye	*1	fam	*3	*0	*1	*1	*0	77
*0	eye	*1	fam	*3	*0	*1	*3	*0	8
*0	eye	*1	fam	*3	*0	*1	*3	*1	8
*0	eye	*1	fam	*3	*0	*1	*4	*0	9
*0	eye	*1	fam	*3	*1	*0	*3	*1	16
*0	eye	*1	fam	*3	*1	*0	*4	*1	44
*0	eye	*1	fam	*3	*1	*1	*1	*1	30
*0	eye	*2	unfam	*1	*0	*0	*1	*0	24
*0	eye	*2	unfam	*1	*0	*0	*1	*1	25
*0	eye	*2	unfam	*1	*0	*0	*3	*0	32
*0	eye	*2	unfam	*1	*0	*0	*4	*0	51
*0	eye	*2	unfam	*1	*0	*1	*1	*1	45
*0	eye	*2	unfam	*1	*0	*1	*2	*0	89
*0	eye	*2	unfam	*1	*0	*1	*4	*0	32
*0	eye	*2	unfam	*1	*1	*0	*1	*1	29
*0	eye	*2	unfam	*1	*1	*0	*3	*1	43
*0	eye	*2	unfam	*1	*1	*1	*1	*1	18

*0	eye	*2	unfam	*1	*1	*1	*2	*1	12
*0	eye	*2	unfam	*1	*1	*1	*4	*1	25
*0	eye	*2	unfam	*2	*0	*0	*3	*0	34
*0	eye	*2	unfam	*2	*0	*1	*3	*0	29
*0	eye	*2	unfam	*2	*0	*1	*4	*1	36
*0	eye	*2	unfam	*2	*1	*0	*2	*1	30
*0	eye	*2	unfam	*3	*0	*0	*1	*0	49
*0	eye	*2	unfam	*3	*0	*0	*2	*0	47
*0	eye	*2	unfam	*3	*0	*0	*2	*1	38
*0	eye	*2	unfam	*3	*0	*0	*4	*1	72
*0	eye	*2	unfam	*3	*0	*1	*1	*0	63
*0	eye	*2	unfam	*3	*0	*1	*2	*1	4
*0	eye	*2	unfam	*3	*0	*1	*3	*0	40
*0	eye	*2	unfam	*3	*0	*1	*3	*1	21
*0	eye	*2	unfam	*3	*0	*1	*4	*0	40
*0	eye	*2	unfam	*3	*1	*0	*4	*1	13
*0	eye	*2	unfam	*3	*1	*1	*3	*1	43
*0	eye	*2	fam	*1	*0	*0	*3	*0	24
*0	eye	*2	fam	*1	*0	*0	*4	*0	30
*0	eye	*2	fam	*1	*0	*1	*1	*1	49
*0	eye	*2	fam	*1	*0	*1	*3	*0	53
*0	eye	*2	fam	*1	*1	*0	*1	*1	70
*0	eye	*2	fam	*1	*1	*0	*2	*1	41
*0	eye	*2	fam	*1	*1	*1	*2	*1	57
*0	eye	*2	fam	*1	*1	*1	*3	*1	28
*0	eye	*2	fam	*2	*0	*0	*2	*1	38
*0	eye	*2	fam	*2	*0	*0	*4	*0	19
*0	eye	*2	fam	*2	*0	*1	*2	*0	79
*0	eye	*2	fam	*2	*0	*1	*2	*1	48
*0	eye	*2	fam	*2	*0	*1	*4	*0	41
*0	eye	*2	fam	*2	*0	*1	*4	*1	30
*0	eye	*2	fam	*2	*1	*1	*4	*1	38
*0	eye	*2	fam	*3	*0	*0	*1	*0	85
*0	eye	*2	fam	*3	*0	*0	*1	*1	26
*0	eye	*2	fam	*3	*0	*0	*2	*0	23
*0	eye	*2	fam	*3	*0	*0	*3	*0	37
*0	eye	*2	fam	*3	*0	*0	*3	*1	35
*0	eye	*2	fam	*3	*0	*0	*4	*1	15
*0	eye	*2	fam	*3	*0	*1	*1	*0	150
*0	eye	*2	fam	*3	*0	*1	*3	*0	32
*0	eye	*2	fam	*3	*0	*1	*3	*1	28
*0	eye	*2	fam	*3	*0	*1	*4	*0	50
*0	eye	*2	fam	*3	*1	*0	*3	*1	19
*0	eye	*2	fam	*3	*1	*0	*4	*1	47
*0	eye	*2	fam	*3	*1	*1	*1	*1	61
*1	noeye	*1	unfam	*1	*0	*0	*1	*0	14
*1	noeye	*1	unfam	*1	*0	*0	*1	*1	5
*1	noeye	*1	unfam	*1	*0	*0	*3	*0	8
*1	noeye	*1	unfam	*1	*0	*0	*4	*0	9
*1	noeye	*1	unfam	*1	*0	*1	*2	*0	12
*1	noeye	*1	unfam	*1	*0	*1	*2	*1	14
*1	noeye	*1	unfam	*1	*0	*1	*4	*0	21
*1	noeye	*1	unfam	*1	*0	*1	*4	*1	7
*1	noeye	*1	unfam	*1	*1	*0	*3	*1	14
*1	noeye	*1	unfam	*1	*1	*1	*1	*1	39
*1	noeye	*1	unfam	*1	*1	*1	*2	*1	9
*1	noeye	*1	unfam	*1	*1	*1	*3	*1	5
*1	noeye	*1	unfam	*2	*0	*0	*2	*0	5
*1	noeye	*1	unfam	*2	*0	*0	*2	*1	9
*1	noeye	*1	unfam	*2	*0	*0	*3	*1	3
*1	noeye	*1	unfam	*2	*0	*0	*4	*0	16
*1	noeye	*1	unfam	*2	*0	*0	*4	*1	3
*1	noeye	*1	unfam	*2	*0	*1	*1	*0	7
*1	noeye	*1	unfam	*2	*0	*1	*1	*1	4
*1	noeye	*1	unfam	*2	*0	*1	*3	*0	21
*1	noeye	*1	unfam	*2	*0	*1	*3	*1	6
*1	noeye	*1	unfam	*2	*0	*1	*4	*0	11
*1	noeye	*1	unfam	*2	*1	*0	*2	*1	17
*1	noeye	*1	unfam	*2	*1	*0	*4	*1	8
*1	noeye	*1	unfam	*3	*0	*0	*3	*0	5
*1	noeye	*1	unfam	*3	*0	*1	*2	*0	5
*1	noeye	*1	unfam	*3	*1	*0	*1	*1	9
*1	noeye	*1	unfam	*3	*1	*1	*4	*1	2
*1	noeye	*1	fam	*1	*0	*0	*1	*0	15
*1	noeye	*1	fam	*1	*0	*0	*2	*1	9
*1	noeye	*1	fam	*1	*0	*0	*4	*0	9
*1	noeye	*1	fam	*1	*0	*0	*4	*1	2
*1	noeye	*1	fam	*1	*0	*1	*1	*0	2
*1	noeye	*1	fam	*1	*1	*0	*3	*1	1
*1	noeye	*1	fam	*2	*0	*0	*2	*0	17
*1	noeye	*1	fam	*2	*0	*0	*4	*0	10
*1	noeye	*1	fam	*2	*0	*1	*1	*0	6
*1	noeye	*1	fam	*2	*0	*1	*1	*1	29
*1	noeye	*1	fam	*2	*0	*1	*2	*0	18
*1	noeye	*1	fam	*2	*0	*1	*2	*1	14
*1	noeye	*1	fam	*2	*0	*1	*3	*0	31
*1	noeye	*1	fam	*2	*0	*1	*3	*1	4
*1	noeye	*1	fam	*2	*0	*1	*4	*0	50
*1	noeye	*1	fam	*2	*1	*0	*2	*1	6
*1	noeye	*1	fam	*2	*1	*0	*4	*1	12
*1	noeye	*1	fam	*2	*1	*1	*1	*1	13
*1	noeye	*1	fam	*2	*1	*1	*2	*1	11
*1	noeye	*1	fam	*2	*1	*1	*3	*1	5
*1	noeye	*1	fam	*2	*1	*1	*4	*1	11
*1	noeye	*1	fam	*3	*0	*0	*1	*0	8
*1	noeye	*1	fam	*3	*0	*0	*1	*1	3
*1	noeye	*1	fam	*3	*0	*0	*3	*0	11
*1	noeye	*1	fam	*3	*0	*0	*3	*1	14
*1	noeye	*1	fam	*3	*0	*1	*2	*0	17
*1	noeye	*1	fam	*3	*0	*1	*4	*1	3
*1	noeye	*1	fam	*3	*1	*0	*1	*1	16
*1	noeye	*2	unfam	*1	*0	*0	*1	*0	28
*1	noeye	*2	unfam	*1	*0	*0	*1	*1	7

*1	noeye	*2	unfam	*1	*0	*0	*3	*0	3
*1	noeye	*2	unfam	*1	*0	*0	*4	*0	4
*1	noeye	*2	unfam	*1	*0	*1	*1	*0	5
*1	noeye	*2	unfam	*1	*0	*1	*2	*0	21
*1	noeye	*2	unfam	*1	*0	*1	*2	*1	9
*1	noeye	*2	unfam	*1	*0	*1	*4	*0	28
*1	noeye	*2	unfam	*1	*0	*1	*4	*1	10
*1	noeye	*2	unfam	*1	*1	*0	*3	*1	18
*1	noeye	*2	unfam	*1	*1	*1	*1	*1	60
*1	noeye	*2	unfam	*1	*1	*1	*2	*1	11
*1	noeye	*2	unfam	*1	*1	*1	*3	*1	2
*1	noeye	*2	unfam	*2	*0	*0	*2	*0	18
*1	noeye	*2	unfam	*2	*0	*0	*2	*1	4
*1	noeye	*2	unfam	*2	*0	*0	*3	*1	18
*1	noeye	*2	unfam	*2	*0	*0	*4	*0	15
*1	noeye	*2	unfam	*2	*0	*0	*4	*1	6
*1	noeye	*2	unfam	*2	*0	*1	*1	*0	15
*1	noeye	*2	unfam	*2	*0	*1	*1	*1	10
*1	noeye	*2	unfam	*2	*0	*1	*3	*0	18
*1	noeye	*2	unfam	*2	*0	*1	*3	*1	14
*1	noeye	*2	unfam	*2	*0	*1	*4	*0	1
*1	noeye	*2	unfam	*2	*1	*0	*2	*1	50
*1	noeye	*2	unfam	*2	*1	*0	*4	*1	11
*1	noeye	*2	unfam	*3	*0	*0	*3	*0	4
*1	noeye	*2	unfam	*3	*0	*1	*2	*0	8
*1	noeye	*2	unfam	*3	*1	*0	*1	*1	14
*1	noeye	*2	unfam	*3	*1	*1	*4	*1	11
*1	noeye	*2	fam	*1	*0	*0	*1	*0	2
*1	noeye	*2	fam	*1	*0	*0	*2	*1	27
*1	noeye	*2	fam	*1	*0	*0	*4	*0	11
*1	noeye	*2	fam	*1	*0	*0	*4	*1	10
*1	noeye	*2	fam	*1	*0	*1	*1	*0	14
*1	noeye	*2	fam	*1	*1	*0	*3	*1	2
*1	noeye	*2	fam	*2	*0	*0	*2	*0	40
*1	noeye	*2	fam	*2	*0	*0	*4	*0	13
*1	noeye	*2	fam	*2	*0	*1	*1	*0	2
*1	noeye	*2	fam	*2	*0	*1	*1	*1	37
*1	noeye	*2	fam	*2	*0	*1	*2	*0	14
*1	noeye	*2	fam	*2	*0	*1	*2	*1	19
*1	noeye	*2	fam	*2	*0	*1	*3	*0	39
*1	noeye	*2	fam	*2	*0	*1	*3	*1	10
*1	noeye	*2	fam	*2	*0	*1	*4	*0	75
*1	noeye	*2	fam	*2	*1	*0	*2	*1	23
*1	noeye	*2	fam	*2	*1	*0	*4	*1	25
*1	noeye	*2	fam	*2	*1	*1	*1	*1	15
*1	noeye	*2	fam	*2	*1	*1	*2	*1	7
*1	noeye	*2	fam	*2	*1	*1	*3	*1	15
*1	noeye	*2	fam	*2	*1	*1	*4	*1	16
*1	noeye	*2	fam	*3	*0	*0	*1	*0	25
*1	noeye	*2	fam	*3	*0	*0	*1	*1	16
*1	noeye	*2	fam	*3	*0	*0	*3	*0	25
*1	noeye	*2	fam	*3	*0	*0	*3	*1	7
*1	noeye	*2	fam	*3	*0	*1	*2	*0	16
*1	noeye	*2	fam	*3	*0	*1	*4	*1	26
*1	noeye	*2	fam	*3	*1	*0	*1	*1	38
*1	eye	*1	unfam	*1	*0	*0	*1	*1	7
*1	eye	*1	unfam	*1	*0	*0	*3	*0	11
*1	eye	*1	unfam	*1	*0	*0	*4	*0	6
*1	eye	*1	unfam	*1	*0	*1	*1	*1	5
*1	eye	*1	unfam	*1	*0	*1	*2	*0	19
*1	eye	*1	unfam	*1	*0	*1	*4	*0	1
*1	eye	*1	unfam	*1	*1	*0	*1	*1	5
*1	eye	*1	unfam	*1	*1	*0	*3	*1	10
*1	eye	*1	unfam	*1	*1	*1	*1	*1	1
*1	eye	*1	unfam	*1	*1	*1	*2	*1	1
*1	eye	*1	unfam	*1	*1	*1	*4	*1	3
*1	eye	*1	unfam	*2	*0	*0	*3	*0	4
*1	eye	*1	unfam	*2	*0	*1	*3	*0	11
*1	eye	*1	unfam	*2	*0	*1	*4	*1	11
*1	eye	*1	unfam	*2	*1	*0	*2	*1	4
*1	eye	*1	unfam	*3	*0	*0	*1	*0	12
*1	eye	*1	unfam	*3	*0	*0	*2	*0	10
*1	eye	*1	unfam	*3	*0	*0	*2	*1	3
*1	eye	*1	unfam	*3	*0	*0	*4	*1	16
*1	eye	*1	unfam	*3	*0	*1	*1	*0	6
*1	eye	*1	unfam	*3	*0	*1	*2	*1	1
*1	eye	*1	unfam	*3	*0	*1	*3	*0	6
*1	eye	*1	unfam	*3	*0	*1	*3	*1	7
*1	eye	*1	unfam	*3	*0	*1	*4	*0	10
*1	eye	*1	unfam	*3	*1	*1	*3	*1	6
*1	eye	*1	fam	*1	*0	*0	*3	*0	6
*1	eye	*1	fam	*1	*0	*0	*4	*0	6
*1	eye	*1	fam	*1	*0	*1	*1	*1	11
*1	eye	*1	fam	*1	*0	*1	*3	*0	19
*1	eye	*1	fam	*1	*1	*0	*1	*1	35
*1	eye	*1	fam	*1	*1	*0	*2	*1	7
*1	eye	*1	fam	*1	*1	*1	*2	*1	17
*1	eye	*1	fam	*1	*1	*1	*3	*1	5
*1	eye	*1	fam	*2	*0	*0	*2	*1	8
*1	eye	*1	fam	*2	*0	*0	*4	*0	3
*1	eye	*1	fam	*2	*0	*1	*2	*0	21
*1	eye	*1	fam	*2	*0	*1	*2	*1	10
*1	eye	*1	fam	*2	*0	*1	*4	*0	13
*1	eye	*1	fam	*2	*0	*1	*4	*1	10
*1	eye	*1	fam	*2	*1	*1	*4	*1	21
*1	eye	*1	fam	*3	*0	*0	*1	*0	32
*1	eye	*1	fam	*3	*0	*0	*1	*1	10
*1	eye	*1	fam	*3	*0	*0	*2	*0	10
*1	eye	*1	fam	*3	*0	*0	*3	*0	23
*1	eye	*1	fam	*3	*0	*0	*3	*1	6
*1	eye	*1	fam	*3	*0	*0	*4	*1	2
*1	eye	*1	fam	*3	*0	*1	*1	*0	36
*1	eye	*1	fam	*3	*0	*1	*3	*0	9

*1	eye	*1	fam	*3	*0	*1	*3	*1	10
*1	eye	*1	fam	*3	*0	*1	*4	*0	15
*1	eye	*1	fam	*3	*1	*0	*3	*1	11
*1	eye	*1	fam	*3	*1	*0	*4	*1	11
*1	eye	*1	fam	*3	*1	*1	*1	*1	21
*1	eye	*2	unfam	*1	*0	*0	*1	*0	4
*1	eye	*2	unfam	*1	*0	*0	*1	*1	14
*1	eye	*2	unfam	*1	*0	*0	*3	*0	9
*1	eye	*2	unfam	*1	*0	*0	*4	*0	24
*1	eye	*2	unfam	*1	*0	*1	*1	*1	7
*1	eye	*2	unfam	*1	*0	*1	*2	*0	20
*1	eye	*2	unfam	*1	*0	*1	*4	*0	8
*1	eye	*2	unfam	*1	*1	*0	*1	*1	9
*1	eye	*2	unfam	*1	*1	*0	*3	*1	20
*1	eye	*2	unfam	*1	*1	*1	*1	*1	3
*1	eye	*2	unfam	*1	*1	*1	*2	*1	12
*1	eye	*2	unfam	*1	*1	*1	*4	*1	20
*1	eye	*2	unfam	*2	*0	*0	*3	*0	11
*1	eye	*2	unfam	*2	*0	*1	*3	*0	7
*1	eye	*2	unfam	*2	*0	*1	*4	*1	6
*1	eye	*2	unfam	*2	*1	*0	*2	*1	7
*1	eye	*2	unfam	*3	*0	*0	*1	*0	22
*1	eye	*2	unfam	*3	*0	*0	*2	*0	31
*1	eye	*2	unfam	*3	*0	*0	*2	*1	13
*1	eye	*2	unfam	*3	*0	*0	*4	*1	26
*1	eye	*2	unfam	*3	*0	*1	*1	*0	14
*1	eye	*2	unfam	*3	*0	*1	*2	*1	3
*1	eye	*2	unfam	*3	*0	*1	*3	*0	8
*1	eye	*2	unfam	*3	*0	*1	*3	*1	10
*1	eye	*2	unfam	*3	*0	*1	*4	*0	9
*1	eye	*2	unfam	*3	*1	*0	*4	*1	8
*1	eye	*2	unfam	*3	*1	*1	*3	*1	26
*1	eye	*2	fam	*1	*0	*0	*3	*0	11
*1	eye	*2	fam	*1	*0	*0	*4	*0	13
*1	eye	*2	fam	*1	*0	*1	*1	*1	16
*1	eye	*2	fam	*1	*0	*1	*3	*0	19
*1	eye	*2	fam	*1	*1	*0	*1	*1	41
*1	eye	*2	fam	*1	*1	*0	*2	*1	18
*1	eye	*2	fam	*1	*1	*1	*2	*1	24
*1	eye	*2	fam	*1	*1	*1	*3	*1	17
*1	eye	*2	fam	*2	*0	*0	*2	*1	10
*1	eye	*2	fam	*2	*0	*0	*4	*0	8
*1	eye	*2	fam	*2	*0	*1	*2	*0	29
*1	eye	*2	fam	*2	*0	*1	*2	*1	21
*1	eye	*2	fam	*2	*0	*1	*4	*0	26
*1	eye	*2	fam	*2	*0	*1	*4	*1	11
*1	eye	*2	fam	*2	*1	*1	*4	*1	29
*1	eye	*2	fam	*3	*0	*0	*1	*0	60
*1	eye	*2	fam	*3	*0	*0	*1	*1	15
*1	eye	*2	fam	*3	*0	*0	*2	*0	13
*1	eye	*2	fam	*3	*0	*0	*3	*0	27
*1	eye	*2	fam	*3	*0	*0	*3	*1	6
*1	eye	*2	fam	*3	*0	*0	*4	*1	12
*1	eye	*2	fam	*3	*0	*1	*1	*0	34
*1	eye	*2	fam	*3	*0	*1	*3	*0	14
*1	eye	*2	fam	*3	*0	*1	*3	*1	4
*1	eye	*2	fam	*3	*0	*1	*4	*0	7
*1	eye	*2	fam	*3	*1	*0	*3	*1	14
*1	eye	*2	fam	*3	*1	*0	*4	*1	26
*1	eye	*2	fam	*3	*1	*1	*1	*1	23

Appendix C

Significant main effects and interactions of 6 ANOVAs

SOURCE	SUM OF SQUARES	D.F.	MEAN SQUARE	F	TAIL PROB.
game boundary x eyecontact x role x familiarity x sex					
gameboun.	74269244.17945	1	74269244.17945	131.31	0.0000
eye	13400496.44997	1	13400496.44997	23.69	0.0000
role	17030132.41824	1	17030132.41824	30.11	0.0000
fam sex	7619290.61506	2	3809645.30753	6.74	0.0012
eye fam sex	21098299.65524	2	10549149.82762	18.65	0.0000
gameboun. eye fam sex	4930530.54693	2	2465265.27347	4.36	0.0128
3-class-a_move x 3-class-b_move x eyecontact x role					
3-class-b_move	42329879.24814	1	42329879.24814	77.79	0.0000
eye	9257534.23146	1	9257534.23146	17.01	0.0000
role	7438354.40820	1	7438354.40820	13.67	0.0002
eye fam sex	15922601.90168	2	7961300.95084	14.63	0.0000
game boundary x eyecontact x role x match x contrast x route					
gameboun.	89995761.95073	1	89995761.95073	161.95	0.0000
eye	42951349.82110	1	42951349.82110	77.29	0.0000
role	22707960.34805	1	22707960.34805	40.86	0.0000
match	8145704.53547	1	8145704.53547	14.66	0.0001
route	5004193.51608	3	1668064.50536	3.00	0.0292
gameboun. role	2883712.29830	1	2883712.29830	5.19	0.0227
gameboun match	2291204.45445	1	2291204.45445	4.12	0.0423
eye contrast	3617724.89630	1	3617724.89630	6.51	0.0107
role contrast	2485447.17780	1	2485447.17780	4.47	0.0344
eye route	8315803.31872	3	2771934.43957	4.99	0.0018
match route	10408086.58953	3	3469362.19651	6.24	0.0003
contrast route	4502762.50862	3	1500920.83621	2.70	0.0439
gameboun eye contras	3918800.19643	1	3918800.19643	7.05	0.0079
eye role contrast	5612110.67615	1	5612110.67615	10.10	0.0015
gameboun match route	4471009.64909	3	1490336.54970	2.68	0.0451
gameboun contrast route	6362310.05115	3	2120770.01705	3.82	0.0095
eye contrast route	19529557.78611	3	6509852.59537	11.71	0.0000
role contrast route	4519886.11415	3	1506628.70472	2.71	0.0433
match contrast route	7981620.48764	3	2660540.16255	4.79	0.0025
gameboun eye role route	6182586.27035	3	2060862.09012	3.71	0.0111
eye match contrast route	20903450.23512	3	6967816.74504	12.54	0.0000
role match contrast rout	5596822.55632	3	1865607.51877	3.36	0.0180
game eye role match route	4841642.49112	3	1613880.83037	2.90	0.0334
game role match cont route	5409300.47630	3	1803100.15877	3.24	0.0210
eye role match cont route	5474058.81196	3	1824686.27065	3.28	0.0199

game boundary x eyecontact x role x match x task familiarity

gameboun.	93962100.11372	1	93962100.11372	166.09	0.0000
eye	50820812.75749	1	50820812.75749	89.83	0.0000
role	29245328.52454	1	29245328.52454	51.69	0.0000
taskfam.	14813659.02993	1	14813659.02993	26.18	0.0000
match	8820438.27455	1	8820438.27455	15.59	0.0001
eye role	2260425.90781	1	2260425.90781	4.00	0.0456
eye taskfam.	6097567.98804	1	6097567.98804	10.78	0.0010
role taskfam.	4672440.06191	1	4672440.06191	8.26	0.0041
gameboun. match	4440988.64325	1	4440988.64325	7.85	0.0051
game eye taskfam.	3843239.36891	1	3843239.36891	6.79	0.0091
eye role taskfam.	2856976.32363	1	2856976.32363	5.05	0.0246
game taskfam match	3090549.89444	1	3090549.89444	5.46	0.0194

(6 level) a-move x (9 level) b-move

a_move	12758647.53631	5	2551729.50726	4.21	0.0008
b_move	22766120.93325	8	2845765.11666	4.70	0.0000
a_move b_move	48456036.56455	40	1211400.91411	2.00	0.0002

game boundary x eyecontact x role x task fam. x a-turn channel x b-turn channel

game boundary	11355557.70322		111355557.70322	20.15	0.0000
eye	3454170.31432	1	3454170.31432	6.13	0.0133

Appendix D

Tables 1-5 relate to the pilot experiment carried out to test the rhythmic coordination hypothesis, as reported in Chapter 5. They show choices by the exchange type. The Figures in brackets indicate the duration of the between-interval.

Listener Choice

Inter speaker interval categories	"short"	"normal"	"long"
short (120 msec)	9	11	0
original (240 msec)	5	14	1
long (360 msec)	3	16	1
very long (480 msec)	3	15	2

Table 1. A breakdown of choices made for exchange number 1:
A: *...and above that there's an east lake*
B: *Right. OK*
$\chi^2 = 8.86$, df = 6, 0.1, $p < 0.25$

		Listener Choice		
		"short"	"normal"	"long"
Inter-speaker interval categories	short (110 msec)	16	4	0
	original (220 msec)	9	10	1
	long (330 msec)	7	12	1
	very long (440 msec)	2	15	3

Table 2. A breakdown of choices made for exchange 2:
A: *Roman baths. Don't have any Roman baths at all*
B: *Well, when you come round the lake, stick close round the edge*
$\chi^2 = 20.75$, df = 6, $p < 0.005$

		Listener Choice		
		"short"	"normal"	"long"
Inter-speaker interval categories	short (110 msec)	9	10	1
	original (220 msec)	9	11	0
	long (330 msec)	3	11	6
	very long (440 msec)	1	7	12

Table 3. A breakdown of choices made for exchange 3:
A: *Can I just go straight to saloon bar?*
B: *Mark {laugh} no. Just wait*
$\chi^2 = 26.32$, df = 6, $p < 0.005$

		Listener Choice		
		"short"	"normal"	"long"
Inter-speaker interval categories	short (365 msec)	4	14	2
	original (730 msec)	2	17	1
	long (1095 msec)	1	4	15
	very long (1460 msec)	0	2	18

Table 4. A breakdown of choices made for exchange 4:
A: *Pass...some cliffs on the east*
B: *Sandstone cliffs?*
$\chi^2 = 48.14$, df = 6, $p < 0.005$

		Listener Choice		
		"short"	"normal"	"long"
Inter-speaker interval categories	short (400 msec)	3	16	1
	original (800 msec)	1	12	7
	long (1200 msec)	0	8	12
	very long (1600 msec)	1	3	16

Table 5. A breakdown of choices made for exchange 5:
A: *Now, do you have a cattle ranch?*
B: *No*
$\chi^2 = 27.3$, df = 6, $p < 0.005$

Bibliography

Abercrombie, D. (1967). *Elements of General Phonetics.* Edinburgh: Edinburgh University Press.

Albert, E. M. (1972). Culture patterning of speech behavior in Burundi. In J. J. Gumperz & D. H. Hymes (eds.) *Directions in Sociolinguistics.* New York: Holt, Rinehart and Winston. 72 - 105.

Allen, G. D. (1972). The location of rhythmic stress beats in English: an experimental study I + II. *Language and Speech*, **15**, 72-100; 179-195.

Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S. & Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, **34 (4)**, 351-366.

Beattie, G. W. (1983). *Talk: an analysis of speech and non-verbal behaviour in conversation.* Milton Keynes: Open University Press.

Beller, H. K. (1970). Parallel and serial stages in matching. *Journal of Experimental Psychology*, **84**, 213-219.

Bernstein, B. B. (1962). Linguistic codes, hesitation phenomena and intelligence. *Language and Speech,* **5**, 31-46.

Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech*, **8**, 148-158.

Brady, P. T. (1969). A model for generating on-off speech patterns in two-way conversation. *Bell System Technical Journal*, **48**, 2445-2472.

Brown, G., Anderson, A., Yule, G., & Shillcock, R. (1983). *Teaching Talk.* Cambridge: Cambridge University Press.

Brown, P. & Levinson, S. (1987). *Politeness: some universals in language usage.* Cambridge: Cambridge University Press.

Burnage, G. (1990). *CELEX - A Guide for Users.* Nijmegen: Centre for Lexical Information, University of Nijmegen.

Campbell, W. N. (1993). Multi-level timing in speech. *Advanced Telecommunications Research Institute Technical Report.*

Carletta, J., Isard, I., Isard, S., Kowtko, J., Doherty-Sneddon, G., & Anderson, A. (1995). The coding of dialogue structure in a corpus. In J. A. Andernach, S. P. van de Burgt & G. F. van der Hoeven (eds.) *Proceedings of the Ninth Twente Workshop on Language Technology: corpus-based approaches to dialogue modelling*. Universiteit Twente, 1995. 25-34.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Classe, A. (1939). *The Rhythm of English Prose*. Oxford: Basil Blackwell.

Cooper, A. M., Whalen, D. H. and Fowler, C. A. (1986). P-centers are unaffected by phonetic categorisation. *Perception and Psychophysics*, **39**, 187-196.

Cooper, A. M., Whalen, D. H. and Fowler, C. A. (1988). The syllable's rhyme affects its P-center as a unit. *Journal of Phonetics,* **16**, 231-241.

Cooper, L. A. & Shepard, R. N. (1973). Mental rotation of letters. In W. G. Chase (Ed.) *Visual Information Processing*. New York: Academic Press.

Coulthard, M. (1977). *An Introduction to Discourse Analysis*. London: Longman.

Couper-Kuhlen, E. & Auer, P. (1988). On the contextualizing function of speech rhythm in conversation: Question-Answer sequences. *Linguistics Working Group, University of Konstanz. KontRI No. 1.*

Couper-Kuhlen, E. (1993). *English speech rhythm: form and function in everyday verbal interactions.* Amsterdam: John Benjamins.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Darwin, C. J. and Donovan, A. (1980). Perceptual studies of speech rhythm: isochrony and intonation. In: *Spoken Language Generation and Understanding*. Proceedings of the NATO Advanced Study Institute. Dordrecht: D. Reidel.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalysed. *Journal of Phonetics*, **11**, 51-62.

Dauer, R. M. (1987). Phonetic and phonological components of language rhythm. *Proceedings, XIth International Congress of Phonetic Sciences,* **5**, 447-450.

Davidson, J. (1984). Subsequent versions of invitations, offers, requests, and proposals dealing with potential or actual rejection. In J. Maxwell Atkinson & J. Heritage (Eds.) *Structures of social action (studies in emotion and social interaction)* Cambridge: Cambridge University Press, 102-128

De Long , A. J. (1974). Kinesic signals at utterance boundaries in preschool children. *Semiotica,* **11**, 43-73

Donovan, A. and Darwin, C. J. (1979). The perceived rhythm of speech. *Ninth International Congress of Phonetic Sciences*, **1**, 268-274.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversation. *Journal of Personality and Social Psychology*, **23**, 283-292.

Duncan, S. (1973). Toward a grammar for dyadic conversation. *Semiotica*, **9**, 29-46.

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society,* **3**, 161-180.

Duncan, S. & Fiske, D. W. (1977). *Face-to-face Interaction: research, methods, and theory*. Hillsdale, NJ: Lawrence Erlbaum.

Eichelman, W. H. (1970). Familiarity effects in the simultaneous matching task. *Journal of Experimental Psychology*, **86**, 275-282.

Ekman, P. & Friesen, W. V. (1969). The repertoire of non-verbal behavior: categories, origins, usage, and coding. *Semiotica,* **1**, 49-98.

Erickson, F. & Shultz, J. (1982). *The Counselor as Gatekeeper. Social interaction in interviews*. New York: Academic Press.

Ferguson, J. (1975). Interruptions in spontaneous dialogue. Paper delivered to BPS conference, Stirling, 1975.

Firth, J. R. (1935). The technique of semantics. *Papers in Linguistics 1934-1951*, London: Oxford University Press, 1957, 7-33.

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.), *Categories of Human Learning*. New York: Academic Press.

Ford, C. E. & Thompson, S. A. (1995). Interactional Units in Conversation: syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. A. Schegloff & S. A. Thompson (eds.) *Interaction and Grammar.* Cambridge: Cambridge University Press.

Fowler, C. A. (1979). 'Perceptual centers' in speech production and perception. *Perception and Psychophysics*, **25**, 375-386.

Fox, B. A. (1987). Anaphora and the sturcture of discourse. Cambridge: Cambridge University Press.

Fox, R. A. and Lehiste, I. (1987). The effect of vowel quality variations on stress-beat location. *Journal of Phonetics*, **15**, 1-13.

Giegerich, H. J. (1985). *Metrical Phonology and Phonological Structure. German and English.* Cambridge: Cambridge University Press.

Goodwin, C. (1981). *Conversational Organization: interaction between speakers and hearers.* New York: Academic Press.

Goodwin, C. & Goodwin, M. H. (1987). Concurrent operations on talk: notes on the interactive organization of assessments. *IPRA Papers in Pragmatics* **1.1**, 1-54

Halliday, M. A. K. (1985). *An Introduction to Functional Grammar.* London: Edward Arnold.

Hayes, B. (1984). The phonology of rhythm in English. *Linguistic Inquiry*, **15**, 33-74.

Hoequist, C. E. (1983). The perceptual center and rhythm categories. *Language and Speech*, **26**, 367-376.

Huggins, A. W. F. (1972a). Just noticeable differences for segment duration in natural speech. *Journal of the Acoustical Society of America*, **51**, 1270-1278.

Huggins, A. W. F. (1972b). On the perception of temporal phenomena in speech. *Journal of the Acoustical Society of America*, **51**, 1279-1290.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology,* **45**, 188-196.

Jaffe, J. & Feldstein, S. (1970). *Rhythms of Dialog*. New York: Academic.

Jefferson, G. (1972). Side sequences. In *Studies in Social Interaction*. D. Sudnow (Ed.) New York: Free Press. 294-338.

Jefferson, G. (1973). A case of precision timing in ordinary conversation: overlapped tag-positioned address terms in closing sequences. *Semiotica*, **9**, 47-96.

Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a "standard maximum" silence of approximately one second in conversation. In D. Roger and P. Bull (Eds.), *Conversation*, 166-196. Clevedon: Multilingual Matters.

Jones, D. (1956). *An Outline of English Phonetics*. Cambridge: W. Heffer & Sons.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologic,*. **26**, 22-63.

Kowtko, J. C., Isard, S. D., & Doherty-Sneddon, G. M. (1992). Conversational Games within Dialogue. Research Paper HCRC/RP-31. Human Communication Research Centre, Edinburgh.

Labov, W. (1970). The study of language in its social context. *Studium Generale,* **23**, 30-87

Lehiste, I. (1977). Isochrony Reconsidered. *Journal of Phonetics*, **5**, 253-263.

Lerner, G. H. (1987). *Collaborative turn sequences: sentence construction and social action*. Unpublished PhD thesis, University of California, Irvine.

Levelt, W. J. M. (1989). *Speaking; from intention to articulation*. Cambridge, MA: MIT Press.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Lewis, D. K. (1969). Convention: a philosophical study. Cambridge, MA: Harvard University Press.

Liberman, M. (1975). The Intonational System of English. MIT PhD Dissertation. Published 1979. New York: Garland.

Liberman, M. & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, **8**, 249-336.

Local, J. & Kelly, J. (1986). Projection and 'silences': notes on phonetic and conversational structure. *Human Studies*, **9**, 185-204.

Malam, S. (1996). Politeness, accommodation and divergence: implications for sex-difference theory. *Proceedings of the Fifth Manchester Postgraduate Linguistics Conference*, 101-111.

Marcus, S. M. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception and Psychophysics*, **30**, 247-256.

Martin, J.G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review*, **79**, 487-509.

Matarazzo, J. D., Weitman, M., Saslow, G., & Wiens, A. N. (1963). Interviewer influence on durations of interviewee speech. *Journal of Verbal Learning and Verbal Beahavior,* **1**, 451-458.

McClave, E. (1994). Gestural Beats: the rhythm hyopthesis. *Journal of Psycholinguistic Research*, **23**, 45-66.

Mitchell, T. F. (1957). The language of buying and selling in Cyrenaica. *Hesperis,* **44**, 31-71

Morton, J., Marcus, S. and Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review*, **83**, 405-408.

Nespor, M. & Vogel, I. (1986). *Prosodic Phonology*. Dordrecht: Foris.

Nespor, M. & Vogel, I. (1989). On clashes and lapses. *Phonology*, **6**, 69-116.

Oreström, B. (1983). Turn-taking in English conversation. *Lund Studies in English,* **66**. Lund: CWK Gleerup.

Parkman, J. M. & Groen, G. (1971). Temporal aspects of simple additions and comparison. *Journal of Experimental Psychology,* **89**, 335-342.

Pike, K. L. (1945). *The Intonation of American English.* Ann Arbor, Mich.: University of Michigan Publications.

Posner, M. I. (1978). *Chronometric Explorations of Mind.* New York: Oxford University Press.

Reisman, K. (1974). Contrapuntal conversations in an Antiguan village. In *Explanations in the Ethnography of Speaking*. R. Bauman & J. Sherzer (Eds.). Cambridge: Cambridge University Press.

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal (ed.). *Linguistic Controversies, Essays in linguistic theory and practice.* 73-79. London: Edward Arnold.

Rogers, M. G. K. (1974). *Visual and Verbal Processes in the Recognition of Faces.* Unpublished Doctoral dissertation, University of Oregon.

Sacks, H. (1972). On the analysability of stories by children. In J. J. Gumperz & D. Hymes (eds.) *Directions in Sociolinguistics*. New York: Holt, Reinhart and Winston.

Sacks, H., Schegloff, E.A. & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation, *Language*, **50**, 696-735.

Schegloff, E. A. (1972). Notes on a conversational practice: formulating place. In D. Sudnow (ed) *Studies in social interaction*. New York: Free Press. 75-119.

Schegloff, E. (1980). Preliminaries to preliminaries: can I ask you a question? *Sociological Inquiry,* **50**, 104-152.

Schegloff, E. (1982). Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In D. Tannen (ed.) *Analyzing Discourse: text and talk.* Washington, DC: Georgetown University Press. 71-93.

Schegloff, E. (1987). Recycled turn beginnings: a precise repair mechanism in conversation's turn-taking organisation. In G. Button & J. R. Lee (eds.) *Talk and social organization.* Clevedon, England: Multilingual Matters. 70-85.

Schegloff, E. (1988). Discourse as an interactional achievement II: an exercise in conversation analysis. In D. Tannen (ed.) *Linguistics in Context: connecting observation and understanding.* Norwood, NJ: Ablex. 135-158.

Schelling, T. C. (1960). *The Strategy of Conflict.* Cambridge, MA: Harvard University Press.

Schiffrin, D. (1987). *Discourse Markers.* Cambridge: Cambridge University Press.

Schuetze-Coburn, S., Shapley, M., & Weber, E. G. (1992). Units of intonation in discourse: acoustic and auditory analyses in contrast. *Language and Speech*, **34**, 207-234.

Scott, D. R., Isard, S. D. and de Boysson-Bardies, B. (1985). Perceptual Isochrony in English and in French. *Journal of Phonetics*, **13**, 155-162.

Selkirk, E. O. (1984). *Phonology and Syntax: the relation between sound and structure.* Cambridge, Mass.: MIT Press.

Shepard, R. N. & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, **171**, 701-703.

Sinclair, J. & Coulthard, R. M. (1975). *Towards an Analysis of Discourse.* London: Oxford University Press.

Sinclair, J. & Coulthard, M. (1992). Towards an analysis of discourse. In M. Coulthard (ed.) *Advances in Spoken Discourse.* London: Routledge, 1-34.

Stalnaker, R. C. (1978). Assertion. In P. Cole (Ed.), *Syntax and Semantics 9: Pragmatics,* New York: Academic Press. 315-332.

Tannen, D. (1984). *Conversational Style: analyzing talk among friends.* Norwood, NJ: Ablex.

Webb, J. T. (1972). Interview synchrony. An investigation of two speech rate measures in an automated standardized interview. In: Studies in Dyadic Communication, A. W. Siegman and B. Pope (Eds.) New York: Pergamon, 115-133.

Wilson, T. P. & Zimmerman, D. H. (1986). The structure of silence between turns in two-party conversation. *Discourse Processes,* **9**, 375-390.

Winkelman, J. H. & Schmidt, J. (1974). Associative confusions in mental arithmetic. *Journal of Experimental Psychology*, **102**, 734-737.

Yngve, V. H. (1979). On getting a word in edgewise. *Papers from the 6th Regional Meeting, Chicago Linguistic Society*, 567-577.